ABSTRACT

Title of Dissertation: GENERATING AND MEASURING PREDICTIONS IN LANGUAGE PROCESSING

Masato Nakamura
Doctor of Philosophy, 2023

Dissertation Directed by: Professor Colin Phillips
Department of Linguistics

Humans can comprehend utterances quickly, efficiently, and often robustly against noise in the inputs. Researchers have argued that such a remarkable ability is supported by prediction of upcoming inputs. If people use the context to infer what they would hear/see and prepare for likely inputs, they should be able to efficiently process the predicted inputs.

This thesis investigates how contexts can predictively activate lexical representations (lexical pre-activation). I address two different aspects of prediction: (i) how pre-activation is generated using contextual information and stored knowledge, and (ii) how pre-activation is reflected in different measures.

I first assess the linking hypothesis of the speeded cloze task, a measure of pre-activation, through computational simulations. I demonstrate that an earlier model accounts for qualitative patterns of human data but fails to predict quantitative patterns. I argue that a model with an additional but reasonable assumption of lateral inhibition successfully explains these

patterns.

Building on the first study, I demonstrate that pre-activation measures fail to align with each other in cases called argument role reversals, even if the time courses and stimuli are carefully matched. The speeded cloze task shows that "role-appropriate" *serve* in *... which customer the waitress had served* is more strongly pre-activated compared to the "role-inappropriate" *serve* in *... which waitress the customer had served*. On the other hand, the N400 amplitude, which is another pre-activation measure, does not show contrasts between the role-appropriate and inappropriate *serve*. Accounting for such a mismatch between measures in argument role reversals provides insights into whether and how argument roles constrain pre-activation as well as how different measures reflect pre-activation.

Subsequent studies addressed whether pre-activation is sensitive to argument roles or not. Analyses of context-wise variability of role-inappropriate candidates suggest that there are some role-inappropriate pre-activations even in the speeded cloze task. The next study attempts to directly contrast pre-activations of role-appropriate and inappropriate candidates, eliminating the effect of later confounding processes by distributional analyses of reaction times. While one task suggests that role-appropriate candidates are more strongly pre-activated compared to the role-inappropriate candidates, the other task suggests that they have matched pre-activation. Finally, I examine the influence of role-appropriate competitors on role-inappropriate competitors. The analyses of speeded cloze data suggest that N400 amplitudes can be sensitive to argument roles when there are strong role-appropriate competitors. This finding can be explained by general role-insensitivity and partial role-sensitivity in pre-activation processes.

Combined together, these studies suggest that pre-activation processes are generally in-

sensitive to argument roles, but some role-sensitive mechanisms can cause role-sensitivity in pre-activation measures under some circumstances.

GENERATING AND MEASURING PREDICTIONS IN LANGUAGE PROCESSING

by

Masato Nakamura

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:
Professor Colin Phillips, Chair/Advisor
Associate Professor Ellen Lau, Co-Advisor
Associate Professor Naomi Feldman
Professor Philip Resnik
Associate Professor L. Robert Slevc, Dean's Representative

Preface

The work reported in this dissertation is highly collaborative. Chapters 2, 4, 5, and 6 report joint work with Colin Phillips. Chapter 3 reports on joint work with Shota Momma and Colin Phillips. This work has also been supported by the National Science Foundation (Doctoral Dissertation Research Improvement Award 2240434 and National Science Foundation Research Traineeship Award 1449815) and the Office of Naval Research (Multidisciplinary University Research Initiatives N00014-18-1-2670).

Acknowledgments

This work has not been possible without the help of a number of people. First of all, I would like to thank my mentor, Colin Phillips. Colin has been a wonderful mentor and a collaborator throughout my life in Maryland. 5 years with him was a journey of discovering puzzles and trying out everything to solve the puzzles, and we enjoyed the entire process together. One of the many things I learned from him is to stop and think about a puzzle very carefully and deeply. When I had an idea that didn't seem to work, he didn't let me give it up easily, and encouraged me to try harder. In fact, a lot of better ideas came out after thinking about the apparently wrong ideas harder following his advice.

Thank you to my committee members, Ellen Lau, Naomi Feldman, Philip Resnik, and Bob Slevc. Ellen has given me many opportunities to think deeply about language and the brain in her classes and journal reading groups she organized. Her excitement about science has always inspired me and made me excited about science as well. I had never thought that I would become a computational linguist before I met Naomi and Philip. Their classes and discussions with them made me realize how insightful computational modeling is as an approach to understanding the human mind, and modeling has actually become one of my favorite parts of research now. Bob provided me with useful feedback from the perspective of production. I also learned a lot from his class on cognitive science where he was an excellent moderator of discussions.

Thank you to my teaching mentors, Peggy Antonisse and Tonia Bleam. They played crucial roles in my training as a teacher in different ways. Peggy taught the first class I worked as a teaching assistant for. I worked with Tonia for three straight semesters. I would never be able to teach my own class without learning to become a teacher through working with them and explicitly getting advice from them.

Thanks to Kim Kwok, who has kept everything running in the department. Whenever I visited her to ask for help (sometimes in a panic), she magically solved the problems right away. Her friendly personality also made it so fun to chat with her.

Thanks to Rosa Lee, Hanna Muller, Katherine Howitt, London Dixon, and Tal Ness, the students and post-doc who collaborated with me. I was so lucky because there was Rosa, a student who is interested in exactly the same thing as me. But I was even luckier because she was an excellent researcher to collaborate with and a wonderful person to interact with. A lot of ideas came out of discussions with her, and it was simply impossible for me to write this thesis without those discussions. I am grateful to Hanna for being my role model as a graduate student. I always aimed to accomplish half of what Hanna did, and they were exactly the right goals for me. Katherine and London did not only bring new perspectives to the team but also much energy and joy to the team and the entire community. Katherine organized a lot of parties and I enjoyed every single one of them I joined. London's kind and supportive personality has always made it great to work with her. I really appreciate them and Joselyn coming back from Amherst to join my defense and celebrate me. It has always been nice to interact with Tal, but she has been increasingly important in my final year. She has provided me helpful advice as a person who is a few steps ahead of me in the career.

Thanks to Anna DiRienzo, Cassidy Wyatt, and Maddy Keen, who supported my work

as research assistants and lab managers. None of the experiments I conducted were possible without their helps. Especially, thanks to Anna's hard work, I was able to conduct many experiments in the last few months that play crucial roles in this thesis.

Thanks to Shota Momma, my collaborator. Shota has helped me make my ideas better and clearer. He also provided me with a lot of support when I tried to reanalyze and extend his work. Thanks to Adrian Staub and Wing Yee Chow for providing helpful comments to my work and sharing their data and stimuli with me. Many of my studies were inspired by their prior work, and having their data and stimuli enabled many studies of mine that were otherwise impossible.

Thanks to my cohort, Craig Thorburn and Cassidy Henry, also known as mc^2. It was very unfortunate that we were able to see each other much less often since the pandemic, but the year we shared an office together provided me a lot of energy and was a very important part of my life in Maryland.

Thanks to everybody who is or was a part of the Department of Linguistics and the Language Science community in Maryland at any point I was in Maryland. I really appreciate everyone who contributed to the communities and I am so proud of being a part of them. Special thanks to Craig, who did the entire journey in the Ph.D. program with me. It was not an easy journey, but I could feel it less challenging because I knew Craig was going through the same processes with me. Thanks to Hisao and Nima Kurokami for making me feel like I am in my home country whenever I talk with them. The dinner Nima cooked when they invited me for Thanks Giving was the best food I had in America. Thanks to Tyler Knowlton for being my catch-mate. We shared our love for baseball and the time we played catch was really fun and refreshing for me. Thanks to Zach Maher. It was so nice to be a roommate

for an HSP conference with Zach. He was extremely supportive when I was trying to move out from my place, and I wouldn't be able to move out without his help. It was also loved the fun things about Germany he told me, and that made me more excited about my next journey in Germany. Thanks to the psycholinguistics check-in group, both the initial small group with Phoebe Gaston, Nick Huang, Hanna Muller, and Lalitha Balachandran as well as the newer group including Rosa Lee, Xinchi Yu, Katherine Howitt, London Dixon, Allison Dods, Sathvik Nair, Sebastian Mancha, Utku Turk, Tal Ness, and Zach Maher.

Thank you to the professors and advisors in Japan. I wouldn't have even tried to become a linguist without them. I was going to specialize in history unless I had taken classes taught by Masato Kobayashi and Tom Gally in the University of Tokyo, who made me realize how exciting linguistics is. I became more and more excited about psycholinguistics as I worked with Yuki Hirose and Takane Ito, my undergraduate advisors. Reiko Mazuka, who I worked with as a research assistant at RIKEN Brain Science Institute, provided a lot of support when I applied to graduate schools and influenced my decision to come to Maryland, which turned out to be one of the best decisions I have made in my life.

Finally, thanks to my family, Chieko Nakamura, Takashi Nakamura, Yui Kosuge, Shigeru Nakamura, Miyoko Nakamura, Fumiko Matsuda, and Uno. All my journey has been possible because of all their love and support.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ACC	Accusative
CSLP	Chow, Smith, Lau, and Phillips (2016)
EEG	Electroencephalography
ELP	Ehrenhofer, Lau, and Phillips (2019)
ERP	Event-related potential
FT	Finishing time
ICA	Independent Component Analysis
NOM	Nominative
OSV	Object-Subject-Verb
PRES	Present
RT	Reaction time
VWP	Visual world paradigm

1 Introduction

Prediction supports various aspects of human cognition. For example, when someone is playing catch, they do not start moving the glove when the ball actually arrives. Knowledge about physics and experience provide clues for how the ball will fly. They use this information and the trajectory of the ball to generate predictions about where the ball will go and use them to successfully catch it.

Language processing is a part of this broader pattern in human cognition. There are predictable patterns in linguistic inputs just like thrown balls. It follows the grammar instead of physics, and the comprehender's experience tells them how the utterances are likely to unfold. Such knowledge and experience, combined with contexts, provide clues for what is likely to appear next in language comprehension. Ample evidence has shown that comprehenders in fact do not wait for new linguistic inputs and passively process them. Rather, the processing of likely future inputs begins even before they are observed, as we will see later in this chapter. Such a head start in language processing facilitates language processing and may enable remarkably quick and robust processing of language.

How do people carry out such predictions in language processing? In order to answer the question it is important to understand the mapping from people's knowledge and contextual information onto prediction. Importantly, such processes of prediction themselves are not observable. Therefore, researchers use observable dependent measures of prediction to indirectly measure it. In order to investigate prediction, it is necessary to have clear and supported theories for the measures. Thus the investigation of prediction can be described as constructing theories about the mechanisms of prediction generation, which is the mapping

from contexts and knowledge to prediction, and also the processes underlying measures of prediction, which is the mapping from prediction to observable measures.

This thesis will focus on the prediction of lexical items in language processing, and aims to develop a consistent account of generation and measurement of lexical prediction. I take "argument role reversals" such as (1a) and (1b) as useful test cases. In prior studies, some measures of prediction were matched for the *served* in two contexts even though it should only be predicted in (1a) but not in (1b). Based on these findings researchers have argued that the roles of arguments *customer* and *waitress* do not influence predictions about the verb, and hence *served* is predicted similarly in the two contexts. On the other hand, other measures were not matched for the two *served*, suggesting that argument roles are appropriately used for prediction and *served* in (1a) is more predicted. Such cases of misalignment can be especially useful because they yield insights about both the mechanism of prediction generation and the theories of measures. By reconciling the apparently contradictory findings and being informed by different measures, we can form a better understanding of whether and how argument roles impact prediction. Furthermore, careful comparison of different measures and examination of the causes of the discrepancy are informative about what different measures reflect. Therefore, in this thesis, I aim to better understand prediction generation and prediction measures through argument role reversals, combining experiments and computational modeling.

(1) a. The restaurant owner forgot which customer the waitress had served.

b. The restaurant owner forgot which waitress the customer had served.

1.1 Prediction in language processing

The process of retrieving lexical information from long-term memory has been considered as parallel activation of lexical representations. In bottom-up lexical processing, given form information such as sound or strings of letters, lexical representations consistent with the observed form are activated in parallel. Retrieval of lexical representations requires a sufficient amount of activation (e.g., Morton, 1969; McClelland & Elman, 1986; Marslen-Wilson, 1990).

The form is not the only source of lexical activation. Lexical representations (and/or semantic features associated with them) can be activated predictively via contextual information as well, even with the absence of the bottom-up information (Federmeier & Kutas, 1999; Lau, Phillips, & Poeppel, 2008). For example, in (2), the context *The cop arrested the ...* provides clues for what will be mentioned next. Therefore *thief* is activated beforehand, which facilitates the processing of the lexical item once it is actually observed. In this thesis, I will henceforth refer to such context-driven lexical activation as "(lexical) pre-activation" for simplicity, following the prior literature.

(2) The cop arrested the thief.

The notion of lexical representations above involves much simplification since lexical representations consist of multiple levels. Levelt (1989) argued that there are three levels: lexeme, lemma, and (lexical) concept. Lexemes represent the forms, lemmas represent the syntactic information, and the concept represents the meanings. The distinctions are important because there are not necessarily one-to-one mappings between the representations in different levels. For example, homophones such as *bank* as a rising ground or as a finan-

cial institution have separate lemmas and concepts, but they share the lexemes. On the other hand, synonyms such as *save* and *rescue* have shared or very similar concepts but have distinct lemmas and lexemes.

Since this study is an initial attempt to bring multiple pre-activation measures into a single model, I will assume in this thesis for simplicity that there is only a single level of lexical representation without making the distinctions between different levels, despite the complexity in the structure of lexical representations. This simplification can be valid under the assumption that activations and/or predictions at one level of lexical representations transparently influence activations/predictions at other levels, as assumed in many models addressing lexical and lexico-semantic processing (interactive activation and competition models: e.g. Chen & Mirman, 2012; Bayesian hierarchical generative models: e.g. Kuperberg & Jaeger, 2016). For example, when the conceptual representations are predicted, the lemmas connected with them are also predicted/activated, and vice versa. However, it is yet to be known how exactly the different levels of lexical representations are connected with each other, but it is left for future investigation.

Many studies have provided evidence that people in fact pre-activate lexical representations using contextual information. Some evidence comes from anticipatory looks at visual objects. A number of studies showed that, when presented with a picture visually and a sentence auditorily, people looked at visual objects that were likely as continuations for the sentences even before they heard the object being mentioned (e.g., Altmann & Kamide, 1999; Kamide, Altmann, & Haywood, 2003). Another type of evidence comes from facilitated processing of more predictable lexical items compared to less predictable ones. It is known that in eye-tracking experiments people spend a shorter time reading predictable lexical items and

skipped them more often compared to less predictable lexical items, which is often considered to reflect lexical processing (e.g. Ehrlich & Rayner, 1981; see Staub, 2015 for review). Furthermore, a number of EEG studies have shown that less predictable lexical items compared to more predictable lexical items induce a reduced neural activity pattern which is considered to reflect the smaller effort for lexical processing (N400: e.g., Kutas & Hillyard, 1984; Kutas & Federmeier, 2011; Nieuwland et al., 2020). The linking hypotheses of these measures are discussed in Section 1.2.2.

1.2 The simple model of lexical pre-activation

In order to advance our investigation of prediction incorporating the theory of lexical prediction generation and prediction measures, I propose a model that I will refer to as the Simple Activation Model as a starting point (Figure 1). The model consists of a continuously varying activation mechanism that is assumed to underlie diverse experimental measures, plus mechanisms that link the shared activation mechanism to specific experimental measures. The key feature of the model is that this brings theories of prediction generation and theories of different pre-activation measures into a single model with a shared underlying currency. I suspect that this model is an oversimplification of the relevant cognitive mechanisms, but it is a valuable tool because it can form the basis of experimental and modeling studies exploring the relationship between different experimental measures. This thesis uses this Simple Activation Model to guide the studies and aims to understand the underlying processes of prediction generation and measurement through elaborating this model. In the rest of this subsection, I synthesize prior studies to describe the features of the Simple Activation Model.

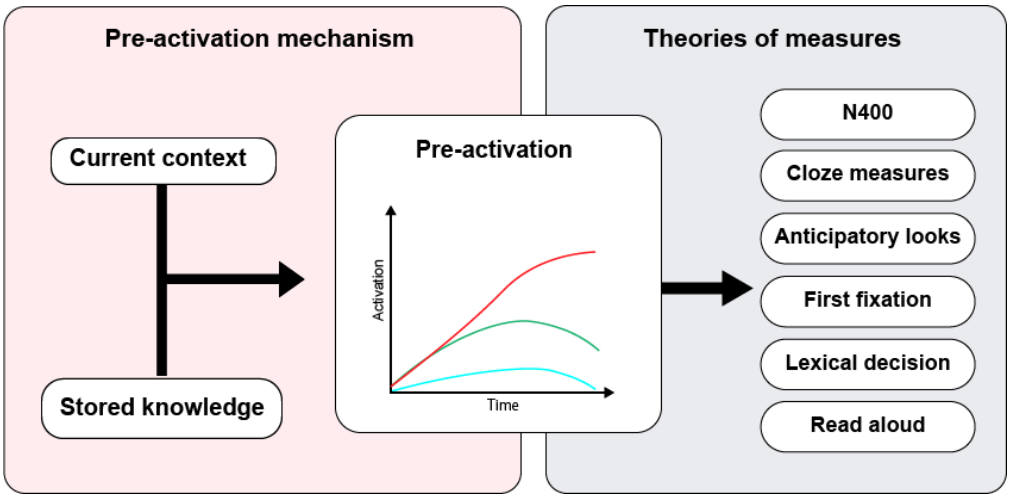


Figure 1: An illustration of the Simple Activation Model. Based on stored knowledge and the context, different lexical items are activated in parallel as possible continuations. Dependent measures reflect the same pre-activation in their own ways.

1.2.1 Pre-activation mechanism

The core feature of the pre-activation mechanism in the Simple Activation Model is to map the contextual information and stored knowledge onto the pre-activation of lexical representations. Some parts of this process can be framed as memory retrieval processes. Some researchers (Chow, Momma, Smith, Lau, & Phillips, 2016; Brouwer, Crocker, Venhuizen, & Hoeks, 2017) argue that this is a memory retrieval process where relevant stored knowledge is retrieved using contextual information. For example, under Chow, Momma, et al. (2016)'s account, in the context of (3), people use the action of *arrest* and the event participant *cop* and activate the likely event representations involving these two that are going to be described. Thus the event of "a cop arresting a thief" or "a cop arresting a suspect" might be activated, and based on the activation of the event representations, the lexical representations associated with the likely event participants in the event (*thief* and *suspect*) are activated.

(3) A cop arrested a ...

Input to the pre-activation mechanism

Contextual information is one of the inputs to the pre-activation mechanism. Many researchers have argued that pre-activation is sensitive to various types of contextual information (simple lexical association: Liao, Lau, & Chow, 2022; verbs' selectional restrictions: Altmann & Kamide, 1999; negation: Nieuwland & Kuperberg, 2008; event knowledge: Kamide et al., 2003; Bicknell, Elman, Hare, McRae, & Kutas, 2010; Rabs, Delogu, Drenhaus, & Crocker, 2022; the relationship between multiple events mediated by discourse connectives: Xiang & Kuperberg, 2015; discourse specific characteristics of objects: Nieuwland & Van Berkum, 2006). The abundant evidence for appropriate activation has led some researchers to argue that people immediately use all available information to activate upcoming lexical items (e.g., Altmann & Mirković, 2009).

The other input is the knowledge of the comprehender. Much fewer studies have been conducted on the structure of the knowledge used for prediction, but there are some prior studies on the models of event knowledge (e.g., Venhuizen, Crocker, & Brouwer, 2019). I will discuss some specific features of the event knowledge representations in Chapters 3 and 7.

Competitive dynamics

One crucial feature of the pre-activation mechanism is that lexical candidates are activated in parallel. For example, when presented with a context (2), people do not only activate the presumably most likely continuation *thief*, but also less likely continuations such as *smuggler* at the same time. The parallel activation assumption is supported by studies showing spreading

activation between candidates. Some studies have shown that unpredictable candidates can be pre-activated by spreading activation from highly-pre-activated lexical candidates that are related to them (e.g., Kleiman, 1980; Kutas & Hillyard, 1984; Federmeier & Kutas, 1999). Such spreading activation requires simultaneous pre-activation of multiple candidates.

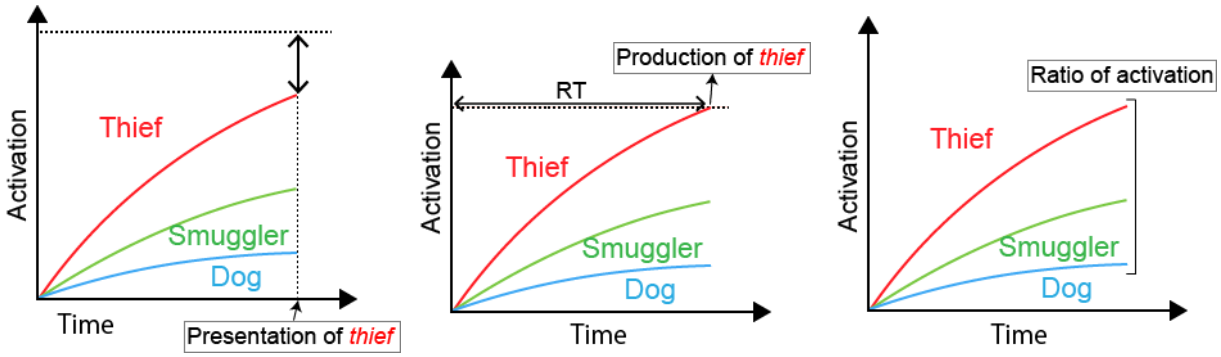
The parallel activation is also supported by studies using a task called the speeded cloze task. The response latency patterns in this task can be straightforwardly explained by competition between lexical items activated in parallel, which will be discussed in Section 1.2.2.

Another important feature is that accumulation of pre-activation is not deterministic but there is some random variability. Even the same lexical item in the same context may accumulate activation differently in different trials. Therefore, while highly predictable lexical items may typically accumulate activations quickly, they may accumulate activations slower than they usually do in some trials. This assumption is necessary to account for the probabilistic responses in the cloze task, which will be introduced in Section 1.2.2 and be discussed in more detail in Chapter 2.

1.2.2 Linking hypotheses

In this model, different experimental measures of prediction are considered to be relatively straightforward reflections of lexical pre-activation. The crucial assumption is that all the different measures take (the same) lexical pre-activation as their inputs, and output observable behavioral or neural measures. Figure 2 illustrate the linking hypotheses of the measures, which are described in more detail in this subsection. As a starting point, I assume that the different experimental measures all tap into the same underlying activation of lexical

candidates, and refer to them as pre-activation measures. However, I will explore in the later chapters the possibility that some of the putative pre-activation measures in fact reflect other processes.



(a) Confrontation measures **(b)** Cloze measures **(c)** Anticipatory looks

Figure 2: Hypothetical figures of measures of pre-activation. Confrontation measures reflect distance from lexical access threshold. Cloze responses are the first lexical items that reach the threshold, and cloze RTs reflect the time to reach the threshold. Anticipatory looks in the Visual World Paradigm reflect the ratio of activation.

Confrontation measures

The amount of pre-activation is often measured by presenting continuations for contexts to people and observing reactions to them. While the exact forms of the output measures are variable, they are the same in that they measure reactions to presented continuations, and the reactions is considered to reflect the ease or difficulty of retrieving lexical representations or the semantic representations associated with them given the presented form of the lexical items. Highly predictable lexical items are easily retrieved because a large amount of activation has accumulated prior to the presentation of the lexical item compared to less predictable items (Figure 2a). These highly pre-activated items can easily reach a threshold for lexical retrieval once they are presented, while it is more difficult for less pre-activated items to reach the same activation threshold. Thus reactions to continuations reveal the distance

from the activation threshold to the accumulated pre-activation. I will refer to these measures as "confrontation measures" of pre-activation because they involve the confrontation of specific continuations and measure reactions to them.

One of the most commonly used confrontation measures is the N400 component. N400 is a component of the event-related potential (ERP), a negative deflection of the ERP starting around 250ms and peaking around 400ms after the onset of a lexical item. Every meaningful lexical item evokes this neural pattern, but it is known that the amplitude of this component follows the predictability of the item given the context. A lexical item that is more predictable in the context elicits smaller N400 amplitude compared to less predictable items (e.g., Kutas & Hillyard, 1984). The increase in N400 amplitude in response to less predictable lexical items is known as the N400 effect (Figure 3). Therefore the amplitude of this component is often interpreted to reflect the amount of activation required to access the lexical item or its meaning, and highly pre-activated lexical items need less of additional activation (Kutas & Federmeier, 2000; Lau et al., 2008). Under this account, if a pair of lexical items elicit a contrast in N400 amplitudes, the difference corresponds to a contrast in the amount of pre-activation that accumulated by the time the items were perceived and processed (Figure 2a).

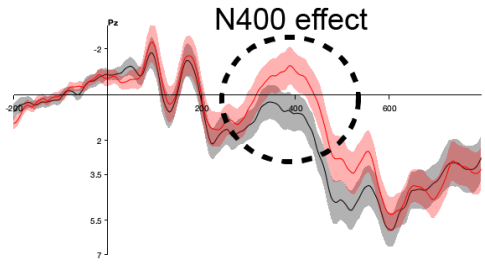


Figure 3: An example of the N400 component and the N400 effect, taken from an EEG experiment in Chapter 4. Negative voltages are plotted upwards.

While I assume here that the N400 component reflects lexical pre-activation, it is important to note that there are competing linking hypotheses regarding the interpretation of N400 amplitude. While the account introduced above claims that the N400 component is sensitive to the activation at the lexical level that accumulated before the bottom-up processing of the lexical item, some researchers have argued that it rather reflects processes to integrating lexical items into combinatorial semantic representations after retrieving lexical representations (e.g., Brown & Hagoort, 1993). Under this integration account, the N400 effect is greater for unexpected items because they tend to be inconsistent with the meanings of the preceding contexts, and hence they are more difficult to integrate into the semantic representation of the sentence. This account differs from the predictive/lexical account of N400 effects in that (i) the N400 reflects the state of the combinatorial representation rather than individual lexical or lexico-semantic representations, and (ii) the process reflected in the N400 is not predictive, in that it occurs after the bottom-up processing of the lexical item.

There is also an account of the N400 component that it reflects both predictive and combinatorial processes. Rabovsky, Hansen, and McClelland (2018) argued that the N400 amplitude is a measure of the cost of updating probabilistic representations of the meaning of the sentence. In their account, people consistently predict what the meaning of an utterance would be as they comprehend its parts incrementally. A highly predictable lexical item requires a minimal update in the predictive utterance meaning representation. On the other hand, unexpected lexical items call for a large update in the semantic representation. I will discuss these possibilities in Chapters 3, 6, and 7.

Although I do not discuss these measures extensively in this thesis, eye movements during reading are also often used as measures of pre-activation. First fixation times and the skipping

rates are often considered to reflect lexical access. The first fixation time is the duration of the first eye fixation on a word until the eyes move to another place, and the skipping rate is the probability of a word not being fixated at all. First fixations on unpredictable words are longer than for predictable words, and more predictable words are more likely to be skipped (Ehrlich & Rayner, 1981; Staub, 2015). These can also be interpreted to reflect the accumulated pre-activation. More pre-activation results in faster completion of lexical retrieval and hence triggers an earlier eye movement to another word. In extreme cases, the pre-activation for a lexical representation is so high that early lexical processing is completed even before people move their eyes on the corresponding word. In such cases, the word is skipped without being fixated on at all (e.g. E-Z Reader model of eye-movement control: Reichle, Pollatsek, & Rayner, 2012).

Other measures of lexical processing can also be used as measures of pre-activation. The lexical decision task is often used to reveal lexical activation. In this task, experiment participants are asked to judge if a presented string of letters is a real word that exists in the target language or is a meaningless non-word. More activated lexical items are faster to be judged as a real word. For example, when a semantically related lexical item is presented prior to the target lexical item ("prime"), lexical decision latencies are known to be shorter, reflecting the spreading activation from the prime (e.g., Meyer & Schvaneveldt, 1971; Perea & Rosa, 2002; Ratcliff, Gomez, & McKoon, 2004). It has also been found that sentential contexts facilitate lexical decision, and that the decision latencies correlate with another measure of pre-activation, the cloze probabilities, which are discussed later in this subsection (Kleiman, 1980).

Finally, the read-aloud task (alternatively, the naming task, or the pronunciation task)

can be used to measure pre-activation through reactions to the presented continuations as well. In this task, people simply read aloud a presented lexical item, and the latency of the speech production onset is measured. It is known that read-aloud latencies are also affected by semantic priming (Stanovich & West, 1983) and by sentential contexts (Seidenberg, Waters, Sanders, & Langer, 1984). Some researchers argue that read-aloud latencies better reflect lexical pre-activations, without the influence of judgment processes, compared to the lexical decision tasks (Seidenberg et al., 1984)

Cloze measures

Pre-activation can also be measured by directly asking people about their predictions without presenting continuations. In the cloze task (Taylor, 1953), participants are presented with sentence fragments and are asked to provide a continuation to each of them. This is different from other measures where the prediction is implicitly probed via the amount of facilitation in the processing of the presented item. Responses are aggregated, and the proportion of responses corresponding to each lexical item is used as a measure of prediction and is called the "cloze probability" of the lexical item. In most cases, it is the number of participants who produced that lexical item divided by the total number of participants. While most cloze studies have collected responses without any time pressure, more recently, some studies have measured the reaction times (RTs) in the cloze task, and used them as a measure of prediction too (e.g., Staub, Grant, Astheimer, & Cohen, 2015; Ness & Meltzer-Asscher, 2021b). Table 1 is a set of hypothetical cloze data following the context (4).

(4) The cop arrested the ...

Item	Cloze probabilities	Cloze RTs
thief	40%	900ms
suspect	20%	1100ms
burglar	10%	1300ms
...	...	
dog	0%	-

Table 1: Hypothetical cloze data for the context (4).

Cloze measures have some potential advantages compared to confrontation measures. Since the task does not involve the direct presentation of continuations, researchers do not need to pre-determine one or a few lexical items to investigate for each context. They can instead analyze the group of lexical items produced by experiment participants. Furthermore, the task allows researchers to avoid the data being skewed by reactions to presented stimuli that are often anomalous.

Staub et al. proposed the Race model as the linking hypothesis for these measures. They argue that, in each trial, the candidate item that reaches the threshold first and wins this "race" is produced. As introduced in the previous section, there is between-trial variability in the speed of activation of the same item in the same context. Therefore, items that are generally activated quickly are more likely to be produced, but sometimes other items can outperform them. Then, the cloze probability is the probability that an item reaches the threshold first, and the RT is the time it takes, on average, to reach the threshold (Figure 2b). Therefore, the items which are activated faster are produced more often and faster, resulting in high cloze probabilities and shorter cloze RTs.

This is in contrast with popular views that preceded this model. Some researchers assumed that cloze production reflects a sampling from the subjective probabilities that people

hold (e.g., Smith & Levy, 2011). In the example above, each participant believes that *thief* is 40% likely to follow the context (4), and *suspect* is 20%, and *burglar* is 10%. They randomly determine which word to produce following these probabilities, and as a result, the cloze probability is roughly identical to the subjective probabilities. While this sampling view and the Race mode agree about the parallel prediction of multiple candidates, there is another view that considers that only a single item is predicted and pre-activated in each trial (e.g., Van Petten & Luka, 2012). The different cloze probabilities of competing candidates arise from individual difference among the participants. In the example above, *thief* is the best continuation for 40% of the participants, while it is not considered at all by the other 60%. Staub et al. (2015) showed that the Race model accounts for the human cloze responses and their RTs the best, as will be reviewed in Chapter 2.

The offline cloze tasks are used much more frequently than the speeded cloze task, but how the offline cloze responses are generated is not known. One simple view is that it reflects the same basic race processes as the speeded cloze task, but there are some additional processes that can modify the responses. For example, in the offline cloze task, participants might compare multiple candidates from a single race that reached the threshold and choose one from those candidates, for example, because the lexical candidate fits the context well, or possibly because that candidate makes them sound smart. Such additional processes are in a black box in that those processes are currently unknown and would also be difficult to model. Since we do not control what participants are doing with the additional time they have in the offline cloze task, those processes could be highly variable between participants or even within participants.

Then, there are two ways to interpret the offline cloze probabilities. One might highlight

the black box component that modifies the race processes. The additional process could be considered to be refining the responses using semi-unlimited time and cognitive resources, and the offline cloze probabilities can be taken to show the upper bound of people's ability to use contextual information in prediction. Most researchers seem to interpret the offline cloze probabilities in this way. In fact, offline cloze probabilities have been mostly used as measures of predictabilities of lexical items given the contexts to norm experiments using other measures.

On the other hand, one might highlight the similarity between the speeded and offline cloze tasks and take the offline cloze probabilities as measures of pre-activation that are noisier than the speeded cloze data due to the black box component. In fact, as Staub et al. (2015) show, there is an extremely high correlation between speeded and offline cloze probabilities (r = .91). This suggests that the contribution of the shared processes is large despite the black box processes. Therefore, the offline cloze probabilities can be interpreted as measures of pre-activation that might be skewed by the additional processes.

Anticipatory looks in the visual world paradigm

Finally, anticipatory eye movements in the visual world paradigm are also commonly used as prediction measures. The visual world paradigm (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995) is a paradigm where participants observe visual objects in front of them as they listen to linguistic inputs auditorily. The choice of the object to look at is considered to reflect real-time lexical activation: people might choose which visual object to fixate on depending on the activation of the lexical representation corresponding to each object. The proportion of looks to each object averaged across trials is considered to reflect the ratio

of lexical activation. In fact, Allopenna, Magnuson, and Tanenhaus (1998) found that a model of lexical activation (TRACE model; McClelland & Elman, 1986), together with the Luce choice rule (R. D. Luce, 1959) to convert activations into response probabilities, closely matched human behavior in the task.

Furthermore, it has been reported that people start looking at visual objects that are likely to be mentioned later in the sentence (e.g., Altmann & Kamide, 1999; Kamide et al., 2003). Therefore the proportion of anticipatory looks at each presented item may reveal the ratio of pre-activation of each item (Figure 2c). This is different from other measures in that it does not measure reactions to continuations that are directly presented as parts of utterances as in confrontation measures, nor does it elicit continuations from participants like the cloze task. Rather, the candidate visual objects have already been presented to participants well before they hear the target lexical item.

1.3 Mismatching measures

The Simple Activation Model predicts that different pre-activation measures should align with each other. This is because of the two assumptions of the Simple Activation Model: (i) the underlying processes generating pre-activation are the same across different measures, and (ii) different measures relatively straightforwardly reflect the common underlying processes. In fact, this is true in most cases, and many prior studies have highlighted the parallelism between different measures, mostly between each measure and (offline) cloze probabilities (e.g., Kleiman, 1980; Seidenberg et al., 1984; Ehrlich & Rayner, 1981; Kutas & Hillyard, 1984; Federmeier, 2007).

However, there are notable exceptions to this pattern. One such exception arises in the

case of argument role reversals. Multiple sources of evidence suggest that comprehenders activate verbs that are inappropriate continuations given the argument roles of prior noun phrases. Concretely, the verb *served* is a likely continuation given the preceding context in (1a), but it is an unlikely continuation in (1b), due to the reversed arguments repeated below as (5a) and (5b). Many studies in different languages reported that the N400 is matched in amplitude for continuations such as *served* in the two contexts, despite the clear contrast in predictability (English: Kuperberg, Sitnikova, Caplan, & Holcomb, 2003; Kim & Osterhout, 2005; Chow, Smith, Lau, & Phillips, 2016; Dutch: Kolk, Chwilla, van Herten, & Oor, 2003; Hoeks, Stowe, & Doedens, 2004; Mandarin: Chow, Lau, Wang, & Phillips, 2018; Liao et al., 2022; Japanese: Oishi & Tsutomu, 2010; However, see Bornkessel-Schlesewsky et al., 2011 for exceptions). There are also some consistent findings from eye-tracking studies (Visual world paradigm: Kukona, Fang, Aicher, Chen, & Magnuson, 2011; eye-tracking while reading: Burnsky, 2022). These findings have been taken as evidence for role-insensitive activation of upcoming lexical items (e.g., Chow, Momma, et al., 2016; Kuperberg, 2016; Brouwer et al., 2017).

(5) a. The restaurant owner forgot which customer the waitress had served.

b. The restaurant owner forgot which waitress the customer had served.

Importantly, we take these studies to inform us about the predictions that people generate based on appropriately identified argument roles, rather than reflecting mis-interpretation. That is, we do not take them to show that people systematically misinterpret the argument role reversal sentences such as (1b) as appropriate sentences such as (1a). In fact, in most of the studies above (Kuperberg et al., 2003; Kolk et al., 2003; Hoeks et al., 2004; Kim &

Osterhout, 2005; Chow et al., 2018; Liao et al., 2022), participants report that argument role reversals such as (1b) are implausible in the comprehension questions. Therefore, even when people obtain the accurate interpretations of the sentences, the N400 component still shows no sensitivity to argument roles. Therefore, we use argument roles to investigate potentially inappropriate predictions, rather than misinterpretations.

On the other hand, people seem to be sensitive to argument roles in the cloze task. Many of the prior studies above used an untimed cloze task to norm the predictability of the target lexical items, and confirmed that role-appropriate items such as *served* in (5a) had clearly greater cloze probabilities compared to role-inappropriate items such as *served* in (5b). Furthermore, Chow, Kurenkov, Kraut, and Phillips (2015) provided suggestive evidence showing that this was also the case for the speeded cloze task. These findings suggest that argument roles are in fact used for pre-activation in contrast with the findings with other measures. Thus different measures of pre-activation tell inconsistent stories about whether argument roles impact lexical pre-activation.

In this thesis, I will mainly explore argument role reversals because they provide useful insights about two interrelated parts of the Simple Activation Model: the pre-activation mechanisms and the theories of measures. First, argument role reversals help us understand the mechanisms generating pre-activation. That is, whether argument roles directly impact the activation profiles, yielding greater activation for role-appropriate lexical candidates and smaller activation for role-inappropriate candidates. In fact, this question is particularly interesting, given that prior N400 and eye-tracking studies are potential exceptions to the general success of human prediction in language comprehension. Considering that people make successful predictions using a variety of contextual information, it is surprising that ar-

gument roles do not seem to influence lexical predictions, even though argument roles provide crucial information about sentence meaning. This raises questions about the apparent pattern that all the available contextual information impacts predictions immediately. Second, the mismatch between measures helps us to refine the linking hypotheses under the Simple Activation Model. As discussed above, each pre-activation measure has its own linking hypothesis motivated by prior studies, and each has been taken to relatively transparently reflect pre-activation. However, this is inconsistent with the finding of a mismatch among the different pre-activation measures. Careful examination of why different measures yield apparently inconsistent findings will help us refine the linking hypothesis of each measure. Such explorations of prediction generation and prediction measures are greatly facilitated by using the Simple Activation Model, which brings multiple measures in a single coherent model.

1.4 Outline of the thesis

This thesis is structured as follows. Before directly examining the discrepancy between pre-activation measures, Chapter 2 investigate the linking hypothesis of the speeded cloze task. As mentioned above, the speeded cloze task has some characteristic advantages compared to other measures. However, fewer studies have been conducted for the processes underlying the speeded cloze task compared to other measures, despite its potential usefulness. In Chapter 2, I first take the Race model proposed by Staub et al. (2015) as the theory of the underlying processes of the speeded cloze task, and argue that the original model cannot explain a quantitative pattern of human speeded cloze data. I propose a refined model keeping the principles of the model, but with the additional assumption of lateral inhibition in pre-

activation. I demonstrate through a computational simulation that this model successfully explains the human speeded cloze data.

Based on this refined linking hypothesis, I subsequently lay out the relationship between underlying cognitive constructs and observable cloze measures in Chapter 2. I demonstrate through a computational simulation that cloze probabilities are measures of levels of activation, whereas cloze reaction times are measures of the rate of activation. Furthermore, I propose a more refined view of the characterization of contexts. Researchers have used the highest cloze probability or the entropy of cloze responses as the standard measure of "contextual constraints" in sentences, where "high constraint" refers to a situation where a context leads to a single lexical item being highly activated while other lexical items receive little activation. I argue that this in fact conflates two different aspects of sets of lexical items that are pre-activated based on a context: (i) the dominance of a single candidate and (ii) the availability of candidates. I demonstrate that different cloze measures can be used to measure each of these.

In Chapter 3, I demonstrate that the N400 and the speeded cloze measures do not align with each other in argument role reversals. I reanalyze a prior Japanese EEG experiment on argument role reversals, and conduct a new speeded cloze experiment that was closely matched with the EEG study in terms of the timing and stimuli. These experiments show that N400 amplitudes are matched for role-appropriate and inappropriate verbs, while cloze responses were predominantly role-appropriate. Such a discrepancy is unexpected given the Simple Activation Model of lexical pre-activation, especially the prior linking hypotheses of the N400 and that of the speeded cloze task I refined in Chapter 2. This can be explained by two competing hypotheses regarding the role of argument roles in pre-activation and

what different measures reflect. The symmetric pre-activation hypothesis maintains that pre-activation may be insensitive to argument roles and N400 reflects that. Under this view, the role-sensitivity found in cloze responses must reflect post-lexical role-sensitive processes. Alternatively, the asymmetric pre-activation hypothesis posits that pre-activation may be sensitive to argument roles, and cloze measures transparently reflect that. Under this approach, N400 amplitudes match for role-appropriate and role-inappropriate verbs in reversal sentences because the N400 is affected by post-lexical role-insensitive processes.

The following chapters examine the symmetric pre-activation hypothesis and the asymmetric pre-activation hypothesis. Chapter 4 shows that role-inappropriate candidates are activated in the speeded cloze task even though they are very rarely produced. A set of speeded cloze experiments are conducted to measure the activation of unspoken role-inappropriate candidates (lures) through the context-wise variability of the strengths of lures. The analyses show that contexts with stronger lures have slower role-appropriate RTs, presumably reflecting the inhibition from the lures. Additionally, contexts with stronger lures are also associated with more frequent production of role-inappropriate responses. These suggest that the role-inappropriate candidates are activated to the extent that they can inhibit role-appropriate responses and can be produced through the race process.

In Chapter 5, I attempt to obtain direct evidence for the symmetric pre-activation hypothesis or for the asymmetric pre-activation hypothesis. I utilized two tasks, the lexical decision task and the read-aloud task, and the distributional analyses of the RTs in those tasks. This allows me to isolate the effect caused by pre-activation from the effect caused by post-lexical processes. The read-aloud task shows no sensitivity to argument roles, but the lexical decision task showed sensitivity to argument roles, which provides mixed evidence for

the role-sensitivity of pre-activation.

Chapter 6 explores the influence of role-appropriate competitors on the role-inappropriate competitors through new analyses of the speeded cloze data collected in Chapter 4. I specifically focus on a finding by Ehrenhofer, Lau, and Phillips (2019), where N400 amplitudes show sensitivity to argument roles when specific stimuli are used. I demonstrate that the role-sensitivity of N400 amplitudes find in Ehrenhofer et al. is caused by strong role-appropriate competitors competing with the role-inappropriate candidates that elicited the N400 effects. I argue that this finding cannot be explained by a generally role-sensitive mechanism with a partial sensitivity to argument roles.

Finally, in Chapter 7, I synthesize the findings from the experiments and computational simulations in this thesis. The contributions from the studies are incorporated into the Simple Activation Model to refine and elaborate the model. I provide refined views on the linking hypotheses of pre-activation measures. I also discuss what pre-activation mechanisms would be able to explain the role-(in)sensitivity of pre-activation found in this study.

2 The speeded cloze task and the Race model

One of the most broadly used measures of lexical prediction is the cloze task (Taylor, 1953). In the cloze task, participants are presented with contexts and they are asked to provide a continuation for each of them. The responses are aggregated and the probability of each response, the cloze probability is obtained. In most cases, it is the number of participants who produced that lexical item in that context divided by the total number of participants. Most cloze studies are untimed and the offline cloze probability is broadly used as a norm for studies using other pre-activation measures, where it is assumed to show the upper limit of prediction given abundant processing resources (e.g., Kutas & Hillyard, 1984). On the other hand, some researchers have recently used the speeded cloze task as an online measure of pre-activation (e.g., Staub et al., 2015; Ness & Meltzer-Asscher, 2021b). This allows us to utilize not only the cloze probability but also the latency of cloze responses, i.e., the cloze reaction times (RTs), which can shed light on different aspects of lexical prediction, as will be discussed in detail in this chapter.

Despite the broad use of the cloze task, surprisingly little is known about the processes underlying the task. Focusing on the speeded cloze task, this study aims to investigate how cloze responses are generated through two sets of computational simulations. While the Race model proposed by Staub et al. (2015) successfully explains the general qualitative patterns of the human speeded cloze data, I demonstrate that it cannot capture the quantitative patterns of the same data set without using unrealistically large values for its parameters. I argue that the quantitative pattern can be successfully explained by maintaining the Race model structures for the underlying processes of the speeded cloze task, but adding one

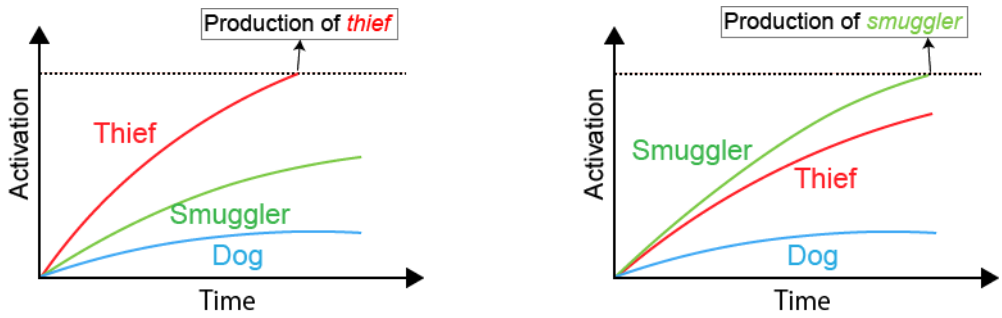
reasonable assumption of lateral inhibitory competition between pre-activated lexical items in the mechanism of prediction generation. This assumption is consistent with findings in bottom-up lexical access.

2.1 The processes underlying the (speeded) cloze task

Staub et al. (2015) proposed a Race model as a theory of the processes underlying the speeded cloze task. This is a member of a class of models called sequential sampling models (see Ratcliff et al., 2016 for review). In sequential sampling models, decision-making processes are understood as a gradual accumulation of evidence. A decision is made once sufficient evidence is gathered and a threshold for evidence is reached. Under the Race model, people activate candidate lexical items in parallel that can follow the context. Different candidates are activated depending on some function of people's long-term knowledge and the context, and the one that reaches the threshold first is produced, just like a race toward a goal. For example, a listener/reader would infer that *thief* is highly likely to follow (6), more than *smuggler*, and therefore it is activated quickly. Then *thief* is likely to reach the threshold first, and this results in more frequent production of *thief* (Figure 4a). The amount of activation shows some variability from trial to trial, and even the same lexical item in the same context might be activated with different rates on different trials. This allows some trials where generally slower candidates such as *smuggler* are activated quickly, and generally more activated ones such as *thief* are activated slowly, resulting in the production of the generally-slower candidate (Figure 4b). Thus the cloze probability reveals the probability of a lexical item reaching the threshold ahead of its competitors (winning a "race"). On the other hand, the cloze RT reflects the time the winner took to reach the threshold, as well as the time

for other processes following the lexical processes such as phonological encoding or motor planning for production. These measures can reveal pre-activation for lexical items because quick pre-activation results in both frequent and fast production.

(6) The cop arrested the ...



(a) A trial where *thief* wins **(b)** A trial where *smuggler* wins

Figure 4: An illustration of two hypothetical trials of races for (6) with different outcomes. *Thief* would win most of the times (e.g. 4a), but *smuggler* can finish before it in some trials (e.g. 4b).

The Race model stands in contrast with popular views that preceded it. Some researchers have proposed that cloze production involves a process of sampling from the subjective probabilities that people hold (e.g., Smith & Levy, 2011). For example, when each participant believes that *thief* is 90% likely to follow the context (6), and *smuggler* is 10%, they randomly determine which lexical item to produce following these probabilities in the cloze task. As a result, the cloze probability is assumed to be roughly identical to the subjective probabilities. While the sampling view and the Race model agree about the parallel activation of multiple candidates, there is another view in which each person only holds a single prediction at a time and only a single item is activated (e.g., Van Petten & Luka, 2012). Under this view, the different cloze probabilities of competing candidates arise from the individual differences among participants or trial-by-trial differences within the same participant. In the example

above, *thief* is the best continuation for 90% of the trials, while it is not considered at all in the other 10%.

The Race model's interpretation of cloze probability might be surprising, especially given the common assumption that the cloze probability of each item, or some kind of probability that is very similar to it, is represented in the human mind (e.g., Smith & Levy, 2011). Rather, under this model, a cloze probability is merely an aggregation of the results of races, and the probability is not itself a part of the cognitive mechanism. What actually exists in the cognitive mechanism is the activation of lexical representations, and the factors that determine how fast each lexical item accumulates activation. The cloze probability is a measure of pre-activation because it reflects these processes, not because the cloze probability itself is directly encoded in the mind.

Staub et al. showed that the Race model in fact captures some key qualitative patterns of the human cloze data they collected in two experiments, while the other theories cannot. One piece of evidence is that cloze probabilities and cloze RTs negatively correlate with each other, such that frequently-produced items ("high-cloze" items) are generally produced faster than rarely-produced items ("low-cloze" items). This is expected under the Race model, because quick activation generally results in both a higher probability of winning a race, which is the high cloze probability, and a shorter time to reach the threshold, which results in the short cloze RT.

An arguably more important piece of support for this model is that it successfully predicts the relationship between a lexical candidate's RTs and the strength of competing lexical candidates. In human data, if two items share the same cloze probability, the item with higher-cloze competitors is likely to be produced faster than the one with lower-cloze

competitors. This can be understood using the Race model. Under the Race model, an item must be activated quickly in order to win races competing with fast competitors. Since high-cloze items are generally activated faster than low-cloze items, an item competing with those high-cloze items is predicted to be produced faster than others.

While the Race mechanism can straightforwardly explain the patterns in the cloze RTs, the other hypotheses fail to make these predictions. If both high- and low-cloze responses are generated by the same sampling process and the high-cloze candidates are simply more likely to be sampled, then there should not be differences between the high-cloze and low-cloze RTs, without any additional assumptions. Similarly, if each person only activates a single candidate on any given trial, and if low-cloze responses are just favored by relatively few people but are still great candidates for them, those low-cloze responses should be produced just as fast as high-cloze responses. The generally long RTs of low-cloze responses suggest that they are not great candidates, even for people who actually produced them, compared to high-cloze responses. Furthermore, it is very difficult for these competing hypotheses to explain why RTs are faster when there is a stronger competitor.

Staub et al. demonstrated the predictions of the Race model through a set of computational simulations, where they simulated the race process as samplings from *finishing time distributions*. A finishing time (FT) is a time for a candidate lexical item to reach the threshold, which is one key component of the RT. The lexical candidate that has the shortest FT in each trial is produced. Because the speed of activation is variable and hence the FT of each item varies from trial to trial, the FT of an item can be considered to follow a probabilistic distribution. Therefore the race process can be modeled as samplings from FT distributions (Figure 5). For example, given context (6), three candidates *thief*, *suspect*, *smuggler* would

have their own finishing time distributions, where the distribution of *thief* may have the smallest mean, which means that it generally has shorter FTs compared to the competitors. If *thief*'s FT is 400ms, *suspect*'s is 800ms, and *smuggler*'s is 1200ms in one trial as in Figure 5, *thief* is produced with an FT of 400ms. On the other hand, if on another trial their FTs are respectively 700ms, 600ms, and 1000ms, *suspect* is then produced with the FT of 600ms.

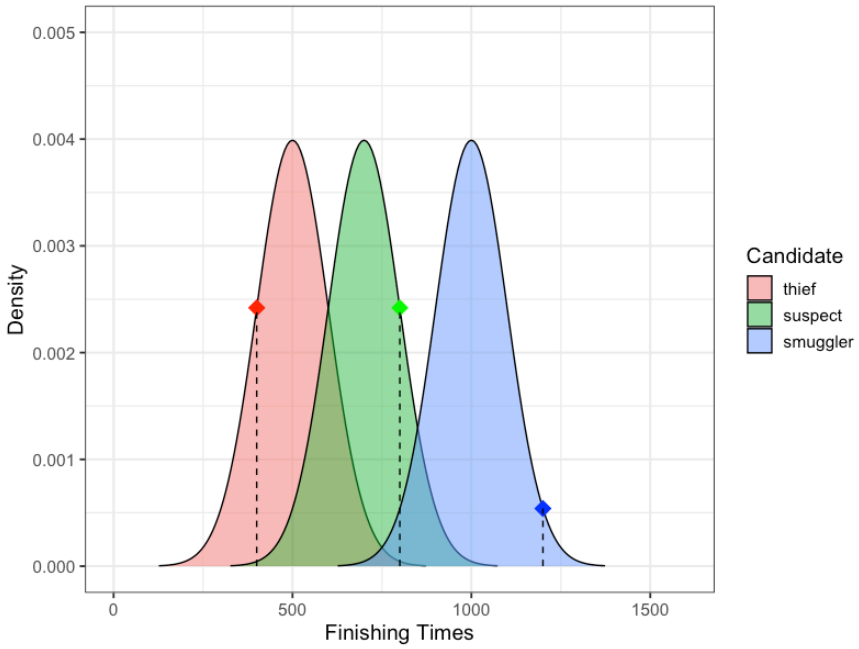


Figure 5: Hypothetical finishing time distributions of *thief*, *suspect*, and *smuggler* given context (6). The three points represent the FTs of those items in one hypothetical trial.

Staub et al. simulated sets of races with different means of FT distributions, and they found that the FT correlated with the cloze probabilities in those simulations (Figure 6). Furthermore, they found that competition with quickly activated items was associated with faster RTs. They used these as support for the model in that the qualitative pattern of the human data was successfully predicted by the model.

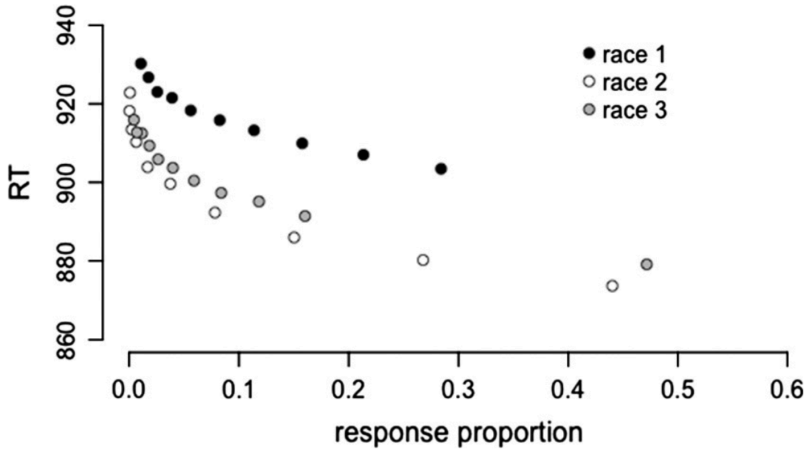


Figure 6: Results of Staub et al. (2015)'s simulation. The three different shades of dots illustrate three simulations. Each simulation consist of 100000 races involving a different set of 10 lexical items. Each dot indicates the response probability (i.e., the cloze probability) of each item and the RT. The negative correlation between the RT and the cloze probability is seen in all simulations.

Thus the model is successful in accounting for the cloze data, but the evidence remains a *qualitative* replication by the simulation. It is not clear how much this model aligns with the *quantitative* pattern of human cloze data. The first aim of this study is to examine whether the slope of the correlation between the cloze probability and the cloze RT is the same in the human data Staub collected. I found that the slopes of correlation are quite different between the human data and the simulated data, because people seem to produce slowly activated items more often than predicted by the Race model. It initially seems that the problem can be solved by adjusting a parameter of the model, but I show that it instead results in an overestimation of the RT variability. Hence, this problem cannot be solved by simply changing the values of the parameters. Secondly, I propose that an additional assumption of lateral inhibition between the predicted candidates can explain such human behavior unexpected by the basic Race model.

2.2 Simulation 1: RT-Probability correlation and FT variability

Although both Staub et al.'s experimental data and their simulated data showed a negative correlation between cloze probability and cloze RTs, juxtaposing the two sets of data, it is clear that there is a large quantitative difference between the correlations (Figure 7). Human data has a much steeper slope compared to the simulations. This means that there are many "slow wins" in human data that are unexpected by the model. As described above, a lexical item must be activated fast enough in order to be produced in a cloze trial so that it reaches the threshold before any other competitors. Therefore, even low-cloze items must be quickly activated when they win. However, in human data, items that we might expect to be too slow to win races are somehow winning with slow FTs, resulting in very long RTs for low-cloze items in human data.

One simple way to address the problem is to change the presumed variability of the FT distributions. If the variability is large, there is more chance that the generally fast competitors happened to be slow, and therefore when they are activated slowly, generally slower items can outperform them. Figure 8 illustrates two races with different FT variabilities. Here, each Lexical item 2 in the two races has the same cloze probability. In Race 1, because Lexical item 1 has a long FT very rarely, Lexical item 2 must be very fast to finish before Lexical item 1. On the other hand, in Race 2, Lexical item 2 can win races without being activated very quickly since the greater variability allows Lexical item 1 to be activated slowly. Thus with greater variability in the FT distributions, generally slow items can win more often with long FTs, allowing more chances of slow wins. This enables the slope in the RT-probability correlation to be steeper because there are more wins with long FTs, resulting

in longer RTs for the low-cloze items.

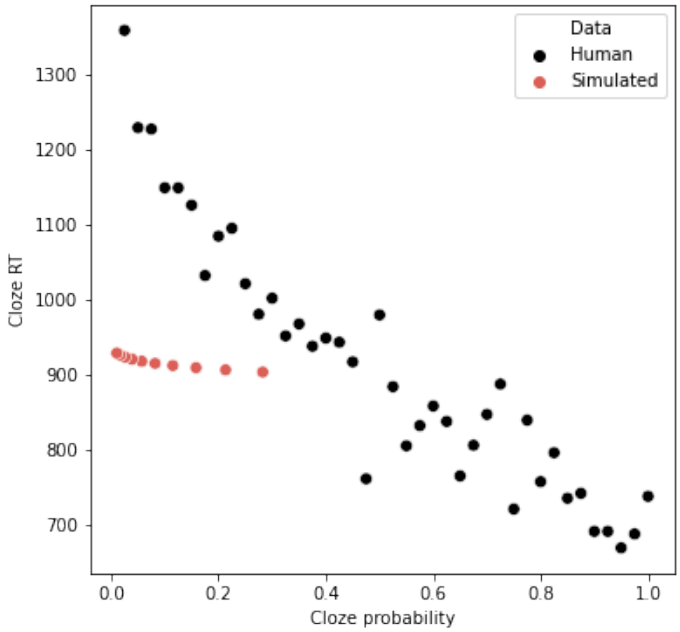


Figure 7: Comparison between the experimental and simulated data in Staub et al. (2015). Each black dot shows the mean RT of all the responses with the same cloze probability across all the different contexts in Staub et al.'s Experiment 2. The red dots show the cloze probability and the RT of each lexical item in Staub et al.'s Race 1 simulation.

In fact, computational simulations with the Race model support this prediction. I ran three sets of races with different variabilities of the FT distributions to examine the effect of FT variability on the cloze RT-probability relationship. Similarly to Staub et al.'s simulations, each race was simulated by randomly sampling FTs of the competitors from their FT distributions, and the item with the shortest FT in each trial is chosen as the winner. However, while Staub et al. just ran races for a single set of competitors, which corresponds to simulating cloze data for a single context, I randomly generated 10,000 sets of 10 competitors for each set of races, just like simulating 10,000 contexts. The FT distributions were

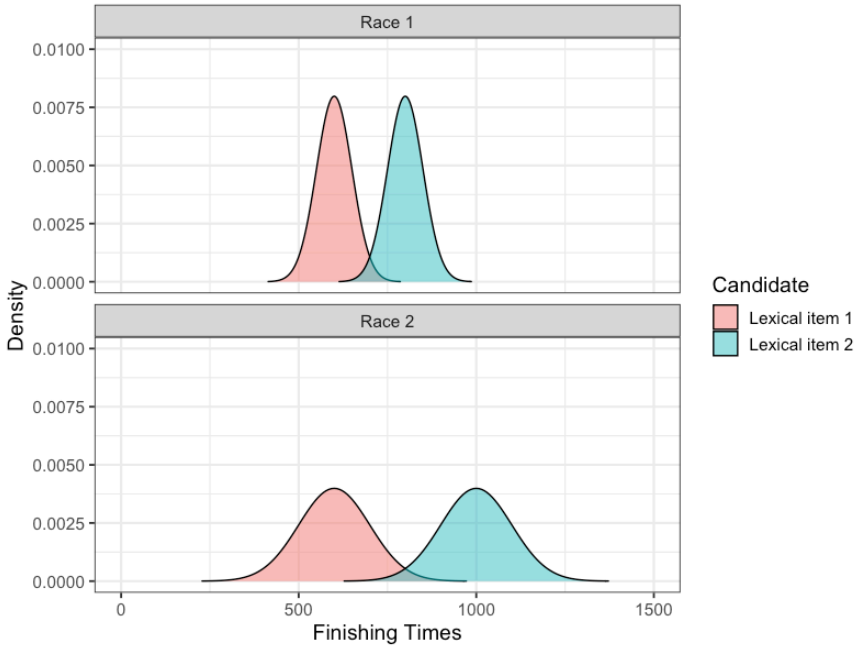


Figure 8: Hypothetical FT distributions with different variability. Each lexical item in the two races has the same cloze probabilities. While Lexical item 2 in Race 1 must have very short FTs to win the race, Lexical item 2 in Race 2 can win with longer FTs.

normal distributions with parameters of μ and σ, where μ corresponds to the mean of the FT distribution and σ corresponds to the variability (standard deviation) of the distribution. Each competitor's FT distribution had a different value for the parameter μ. In Simulation 1, μs of FT distributions were generated from another normal distribution whose parameters μ_μ and σ_μ are listed in the table below. Normal distributions were chosen so that competitive, quickly-activated lexical items are rarer than slower competitors. If the generated μ for the FT distributions was $2.5\sigma_\mu$s further from μ_μ, the values were replaced with the boundary values (either $\mu_\mu + 2.5\sigma_\mu$ or $\mu_\mu - 2.5\sigma_\mu$). For example, when $\mu = 150$ is obtained for Set 1, it is replaced with 250. The purpose of such a replacement is to avoid rare, extremely fast competitors skewing the data. The parameter σ was either 50, 300, or 500, and it was constant within each simulation. I simulated 40 trials for each of the 10,000 contexts, and

obtained the cloze probabilities and RTs, matching Staub et al.'s experimental data, where 40 participants each provided a response. The shortest FT in each race was transformed into an RT by adding 350ms, which is the estimate of the time required to complete post-lexical processes such as motor planning (Indefrey & Levelt, 2004). Figure 9 shows the result of the simulation. As expected, greater variability of the FT distributions results in steeper slopes.

	Set 1	Set 2	Set 3
σ	50	300	500
μ_μ	550	1250	1750
σ_μ	120	350	540

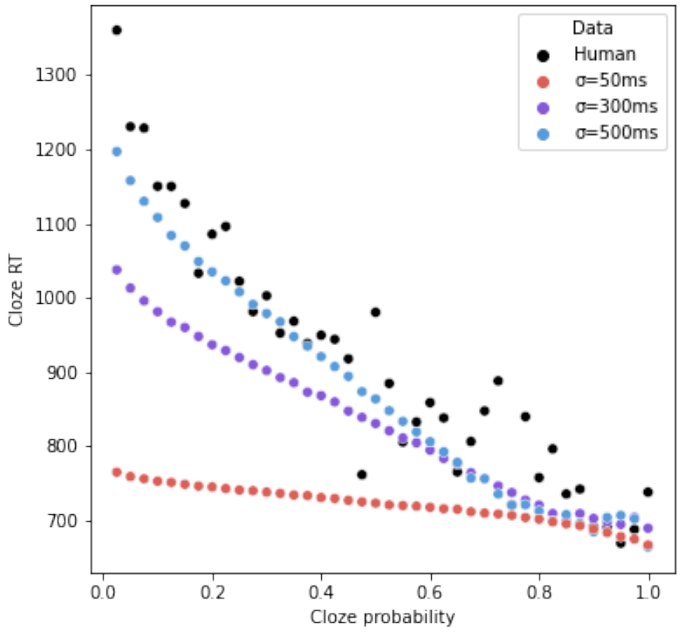


Figure 9: The cloze RT-probability correlation in simulations with different FT-variabilities. Each dot represents the mean RT of all the items with the same cloze probability in each simulation.

I further fit linear models to the data to evaluate the slope in each data set and compared

it with the human data[1]. Since the relationship between the cloze probability and cloze RT is between linear and log-linear (Staub et al., 2015), it is not clear if it is appropriate to consider that RT is a linear function of the cloze probability, or of the logarithm of the cloze probability. Therefore I separately fit RTs against cloze probabilities and log cloze probabilities. Table 2 reports the values of the coefficients. The table shows that the Race model requires about 500ms of FT standard deviation to correctly predict the steep slope in the human RT-probability correlation.

	Coefficient: cloze probability	Coefficient: log cloze probability
Human data	-546.8	-187.1
σ = 50ms	-82.9	-25.0
σ = 300ms	-351.4	-111.8
σ = 500ms	-520.5	-169.0

Table 2: Coefficients of the linear models fit to the human and simulated RT.

While adjusting the parameter of FT varibaility can make the slope of the cloze RT-probability correlation steeper, the variability that I can observe in the human RT data suggests that the FT variability must be much smaller than 500ms. Figure 10 illustrates the variability of the RTs within each lexical item for both human and simulated data sets. I measured how variable RTs are within each lexical item in each context by obtaining the standard deviation of the RTs for each item. (For example, the standard deviation for the RTs of all the *thief* responses for (6)). Then the RT variability was averaged for all the

[1]The linear models are not fit to individual responses, but rather to the mean RT of all responses with the same cloze probability (i.e. the points in Figure 9). Because there are an overwhelmingly large number of singletons in Staub's human data (i.e. 2.5% cloze items), linear models for individual responses would put large weights on those low-cloze items. Since the cloze RT-probability correlation is slightly curved, this would result in an overestimation of the coefficient of the human data. In order to avoid this problem, I fit each model to the mean RT of items with the same cloze probability, so that the singletons would have the same weight on the coefficient as the other items.

lexical items with the same cloze probability. As shown in the figure, greater FT variability results in greater RT variability, and the standard deviation of 500ms for the FT variability (blue dots) results in greater RT variability than the human data (black dots) in the high-cloze responses. Importantly, the simulated RT variability must be much smaller than the RT variability in the human data, because the simulations do not take into account many sources of RT variability other than the FT variability of a lexical item. For example, since RTs are the sum of the FT and the time for other processes such as motor planning for articulation, human RTs must be affected by individual differences, properties of each lexical item, or any other variability in articulatory planning. Therefore, the simulated RTs should be much less variable than the human RTs, and therefore an FT variability of 500ms amounts to a substantial overestimation of the RT variability. Thus increasing the FT variability might explain the slope of the cloze RT-probability correlation, but it causes another issue in the variability of RTs instead.

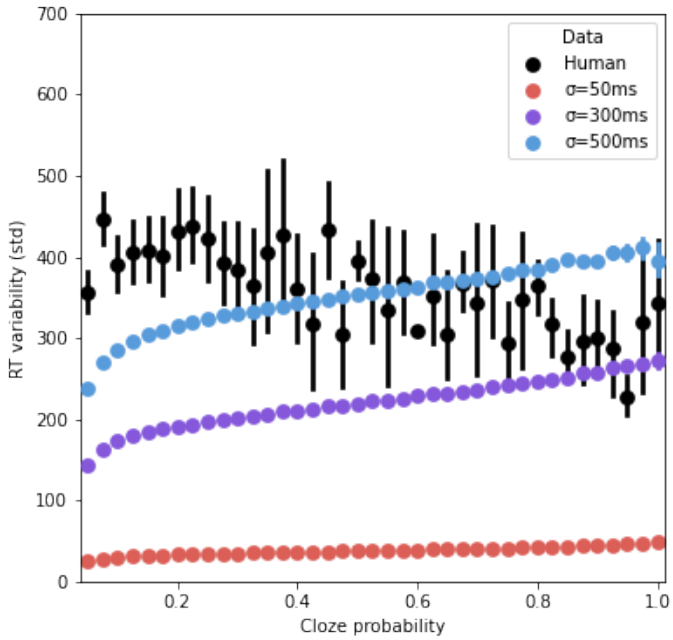


Figure 10: The relationship between the RT variability of each lexical item and the cloze probability. Each point represents the mean of the standard deviation of the RTs within each lexical item.

Thus, the quantitative difference in the RT-probability correlation cannot be explained by simply changing the assumed amount of FT variability, because the amount of variability needed to account for the human RT-probability slope results in implausibly large RT variability. Therefore, the slope of the cloze RT-probability correlation poses a serious problem for the model, which cannot be simply addressed by adjusting the values of the parameters. People seem to often produce slowly activated items that should not be able to win races under the basic version of the Race model.

2.3 Simulation 2: Lateral inhibition

I propose that the Race process underlying the speeded cloze task itself does not need to be revised, but that an additional assumption of lateral inhibition is required to account for the steep RT-probability correlation slope in the human data. The original study by Staub et al. (2015) did not assume any interaction between the competing lexical items, and the activation of each item is determined only by its own properties and by its relationship to the context. However, a common assumption in the lexical access literature is that each lexical item inhibits others, depending on its activation. When an item has a highly activated competitor, its activation slows down or decreases. This feature is shared by many lexical access models such as the TRACE model (McClelland & Elman, 1986). I discuss the support for lateral inhibition in more detail in the discussion section.

Lateral inhibition can explain the large slope of the RT-cloze correlation in the following way. If more pre-activated items inhibit their competitors more strongly, then the winner of a close race must have received greater inhibition from its highly pre-activated competitors. On the other hand, the winner of a non-close race does not have highly activated competitors and hence it is inhibited less. Thus with lateral inhibition, the activation of winners of close races is especially slowed down because of inhibition. Since low-cloze items compete with items that are generally faster than them, they are likely to win primarily in close races where the competitors lost but were highly activated. On the other hand, high-cloze items might sometimes also win close races, but they can also often win races by far because their low-cloze competitors are usually not activated quickly. Therefore, if there is lateral inhibition between competitors, low-cloze items would be more likely to be strongly inhibited

by the competitors than high-cloze items would be. This would result in lengthened FTs and RTs especially for those items. If low-cloze items receive more penalty in their RTs through inhibition, the slope of the RT-cloze correlation would be steeper.

In order to test this hypothesis, I ran a simulation with lateral inhibition. Just like the previous simulations, I ran 40 races with 10 competitors for 1000 contexts. Instead of simulating the races by sampling from the FT distributions, I dynamically modeled the accumulation of activation and inhibition between the competitors. Each race had cycles, and in each cycle, each candidate was both activated and inhibited. The increase in the amount of activation in each cycle followed normal distributions, where each candidate had a different μ (mean) while σ (variability) was uniform across items. The items that had large means would be considered as strong competitors since they accumulate activation quickly. The parameters of the normal distributions were generated from another normal distribution, just like in Simulation 1, with the parameters of $\mu_\mu = 0.02$ and $\mu_\sigma = 0.02$. The amount of inhibition in each cycle was determined by the amount of activation of the competitors. Candidate i inhibits all the other candidates depending on their own activation $a_{i,t}$ at time t, following the formula in (7) (Chen & Mirman, 2012), where the activation $a_{i,t}$ was multiplied by a sigmoid function of the activation. Highly activated items strongly inhibit their competitors, while weakly activated items have small effects (Figure 11), which is independently motivated by neighborhood effects in lexical access studies (Chen & Mirman, 2012). This assumption prevented strongly activated candidates from immediately inhibiting other candidates and allowed multiple candidates to be activated initially. The sigmoid inhibition function (7) in Chen and Mirman (2012) has three parameters. α controls the overall scale of the amount of inhibition. The amount of inhibition suddenly increases at a certain amount of activation

in the function (Figure 11), but the other two parameters x_0 and β are associated with the properties of this steep increase. x_0 corresponds to where the steep increase is on the x-axis. β corresponds to the steepness of the increase. The parameters were mostly adopted from Chen and Mirman ($x_0 = 0.3, \beta = 35$), but the parameter of α was adjusted for the scale of the current simulation ($\alpha = 1.5$). All items had 0.1 activation initially, and the activation was kept to be greater or equal to the lower limit of 0. Following the Race model assumption, the item whose activation first reached the threshold of 1 is produced in each race. The number of cycles was transformed into FTs by assuming that each cycle corresponds to 25ms, and it is further transformed into RTs by adding 350ms, which is the time the post-lexical processes are estimated to take (Indefrey & Levelt, 2004).

(7) $$I_{i,t} = \frac{\alpha a_{i,t}}{1 + e^{-\beta(a_{i,t} - x_0)}}$$

The result shows that the simulation with lateral inhibition closely matches human data (Figure 12), and the slope is similar to the human data (Table 3). Lateral inhibition enabled those slow responses of the low-cloze probability items. Figure 13 illustrates the amount of activation each item lost by inhibition, which is averaged according to the cloze probability of the item. It shows that low-cloze items lost more activation compared to high-cloze items, as predicted. The slight decrease in the amount of lost activation for the lower end of the cloze probability was unexpected. This could be because the extremely low-cloze items sometimes reached the lower limit of activation, where they could not be inhibited anymore. Overall, low-cloze items tend to be slowed down by inhibition more than high-cloze items because

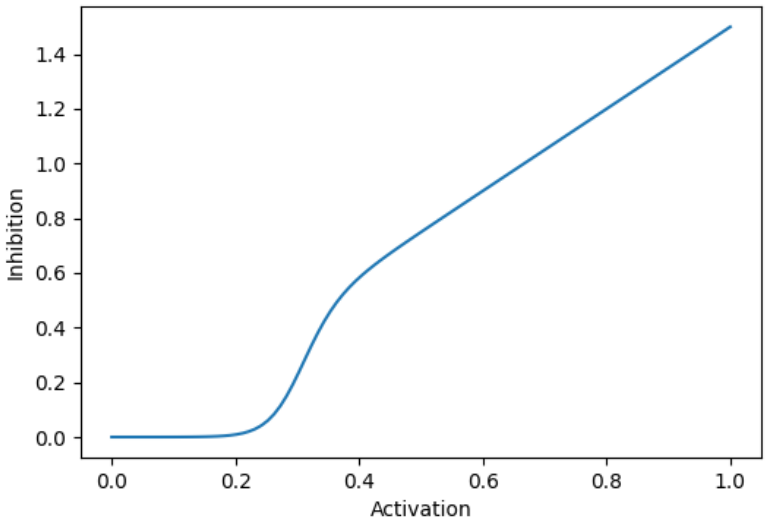


Figure 11: The function of the amount of inhibition each item imposes on its competitors. The parameters in the figure are the same as the simulation.

they tend to win in close races, and therefore the cloze RT-probability slope is much steeper in the model with lateral inhibition.

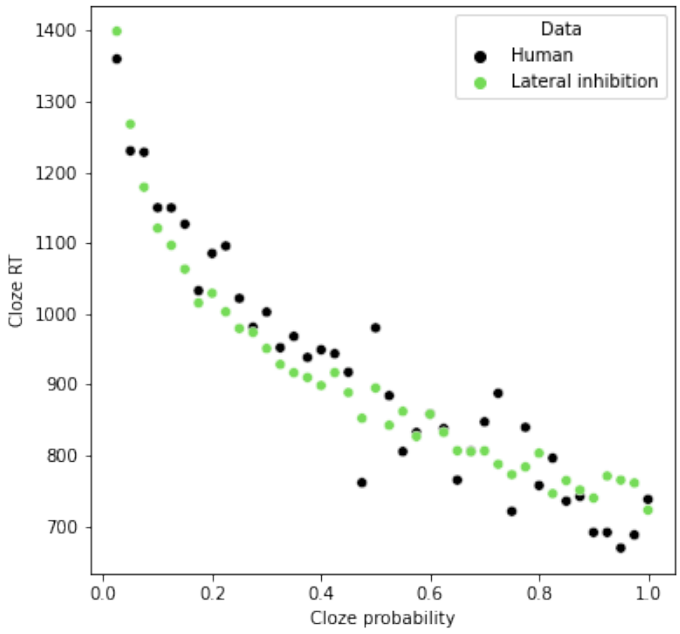


Figure 12: Cloze RT-probability correlation for the human data and simulated data with lateral inhibition.

	Coefficient: cloze probability	Coefficient: log cloze probability
Human data	-546.8	-187.1
Simulated data with inhibition	-473.5	-173.5

Table 3: The coefficients of the linear model of RTs for the human and the simulated data.

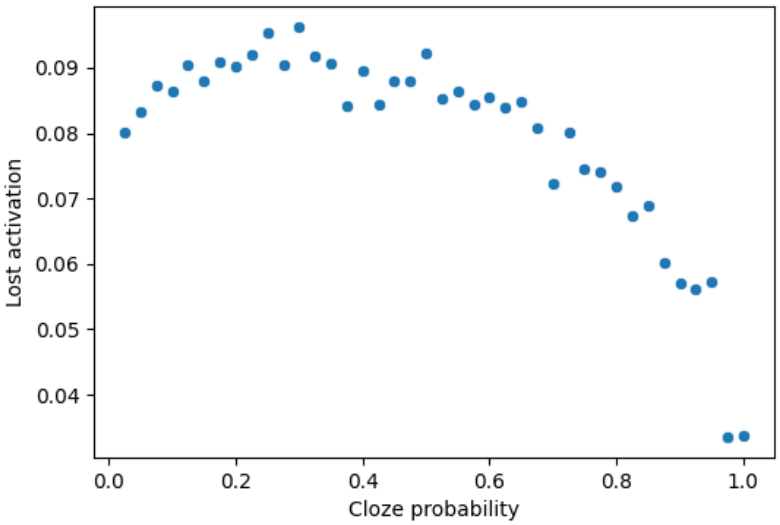


Figure 13: The cloze probability of an item and the amount of activation it lost by inhibition. As predicted, lower-cloze probability items lost more activation by lateral inhibition.

While the RT-probability slope is steeper, the variability of the RT is reduced in this model with lateral inhibition. The model without lateral inhibition results in greater variability in the RTs than human data, whereas the one with lateral inhibition has smaller variability. Figure 14 shows the variability of the FT of each item in the current simulation with the RT variability of human data. It shows that the variability of the simulated FT is consistently below the human RT. Such small variability is plausible considering the numerous factors that make human RTs more variable than simulated RTs, such as post-lexical processes. Thus the Race model with lateral inhibition does not assume too large variability for the FTs, which was the problem of the original Race model with a large FT variability.

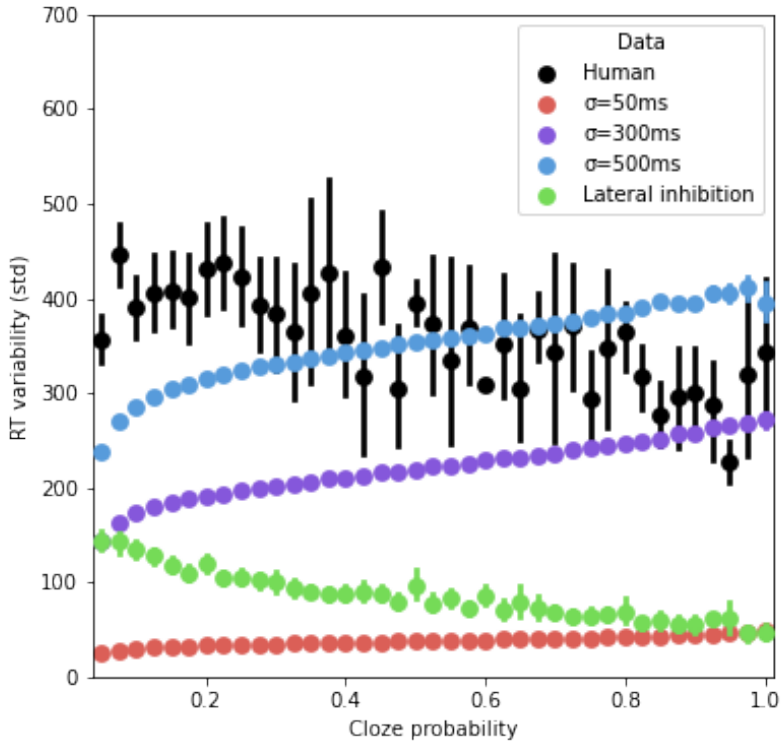


Figure 14: Relationship between the RT variability in the human and simulated data. The red, purple, and blue dots are identical to the dots in Figure 10. The green dots represent the within-lexical item RT variability in the simulation with lateral inhibition.

One concern about lateral inhibition is that its effect might interfere with the relationship between a lexical item's RT and the strength of its competitors. As discussed above, it is found in human cloze data that when two lexical items have matched cloze probabilities, the one that competes with faster competitors is itself produced more quickly, because it has to be activated quickly to outperform the fast competitor. Lateral inhibition might work in the opposite direction, because lexical items competing with faster competitors should be inhibited by them, and hence they should be slower to be produced compared to those without fast competitors.

However, the amount of inhibition needed to account for the slow winners does not cancel the relationship between having short RTs and having fast competitors. In fact, the same model with lateral inhibition also predicts that items that compete with high-cloze items should be produced faster than those that are not. Figure 15 shows the cloze RT-probability correlation for items in high- and low-constraint contexts in the simulation with lateral inhibition. Here the high-constraint contexts are defined as contexts that elicited lexical items with 50% or higher cloze probability while low-constraint contexts do not. The figure shows that high-constraint items are consistently produced quickly compared to low-constraint items across the range of different cloze probabilities. Thus, even with lateral inhibition, the lexical items that compete with faster competitors are still produced faster than those that are not, consistent with Staub et al.'s finding in the human speeded cloze data. Therefore, lateral inhibition between competing lexical items can solve the puzzle of the cloze RT-probability correlation slope, and it can also predict the association between being produced with short RTs and having fast competitors.

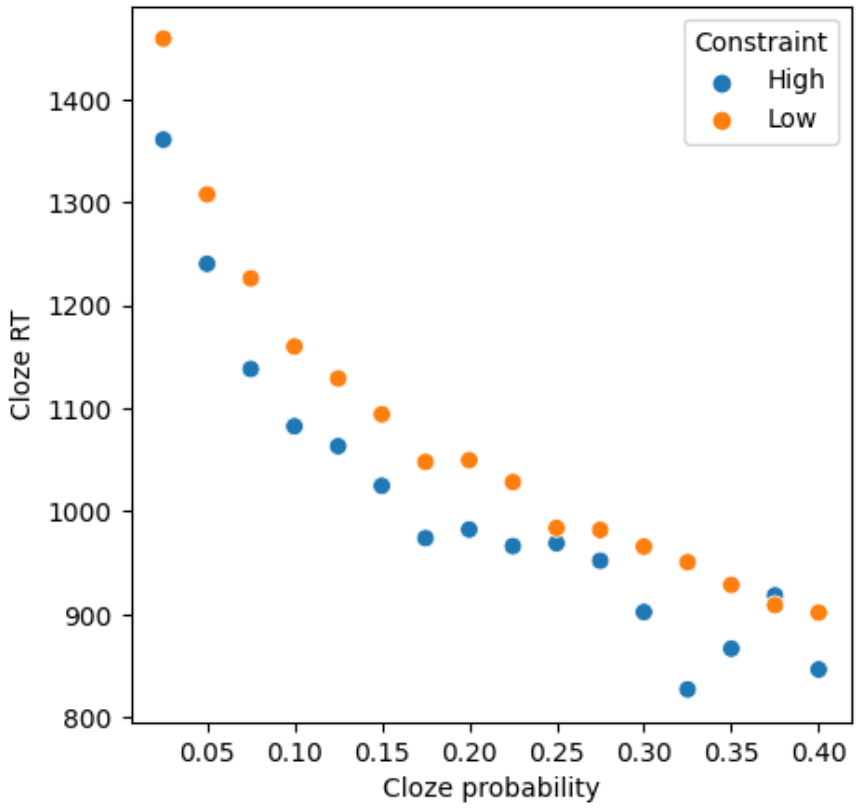


Figure 15: Correlation of cloze RT and cloze probability in high-and low-constraint contexts, where the former has a lexical item with 50% or higher cloze probability.

2.4 Discussion

Despite the importance of lexical predictability in sentence processing and the wide use of the cloze task to operationalize it, there have been surprisingly few studies investigating the cognitive processes underlying the task. As one of such studies, Staub et al. (2015) proposed the Race model where the (speeded) cloze process involves parallel activation of candidate lexical items, and the item that reaches the threshold first is produced as the cloze response. While this model successfully captured the general patterns of the human cloze probability and RTs, only qualitative predictions of the model were examined without detailed discussions

about the quantitative predictions.

The simulations demonstrated that the original Race model without interaction between competing lexical items fails to account for the steep slope in the cloze RT-probability correlation. That is, humans' slow low-cloze responses in the speeded cloze task were unexpected by the model, because those slowly activated items should not have any chance of winning a race. The simulations also showed that those slow wins can be accounted for by changing the parameter of FT variability. However, this adjustment in the FT variability results in an unrealistic amount of RT variability that is greater than the human data. Therefore, the slower low-cloze responses cannot be explained by adjusting this parameter.

A subsequent simulation showed that an additional assumption of lateral inhibition successfully explains this unexpected pattern in the data. Low-cloze items tend to win close races and therefore they are inhibited more than high-cloze ones. This results in the low-cloze responses being much slower than what the model without lateral inhibition expects. Thus these simulations provide evidence that the pre-activation of lexical items involves inhibitory competition between the competing lexical items.

Regarding the Simple Activation Model, which consists of a pre-activation mechanism and the theories of measures, this study has key contributions for both components. First, it provides new insight into the pre-activation mechanism, especially in how pre-activation accumulates. The two computational studies provide support for a lateral inhibition mechanism between competing lexical candidates. Second, it contributes to the theories of measures by reinforcing the Race model as the linking hypothesis for the speeded cloze task. While the correlation between cloze RTs and cloze probabilities was one of the important supports for the Race model, the quantitative pattern was in fact inconsistent with the original Race

model as I demonstrated in Simulation 1. However, I demonstrated in Simulation 2 that the principles of the linking hypothesis do not need to be changed, but just one additional assumption in the process of generating pre-activation can explain the quantitative pattern in human data. Additionally, the implementation of lateral inhibition required a version of Race model with step-by-step activation. This is a significant advance from the original model where the race process was simulated as sampling from FT distributions. Such a step-by-step activation allows for further elaboration of the model, such as a variant with a monitoring mechanism introduced in Chapter 3.

The finding of the current study is consistent with Ness and Meltzer-Asscher (2021b), where lateral inhibition between lexical items contributed to explaining the effect of related and unrelated competitors in a speeded cloze task. They found in their experiment that the activation of a lexical item is more facilitated by a stronger semantically related competitor, but it is more inhibited by a stronger semantically unrelated competitor. Their computational model showed that this is possible when lateral inhibition from strong competitors inhibit the activation of the lexical item, but related competitors can also facilitate the activation through shared semantic features.

Importantly, the findings of this study make the cloze task and pre-activation more consistent and comparable with other cases of lexical activation. Lexical pre-activation can be understood as a special way to access lexical representations without observing the bottom-up information of the lexical item. Prior studies have provided support for lateral inhibition in bottom-up lexical activation. Many studies have shown that highly activated "neighbors" inhibit the processing of the target item. Neighbors of a lexical item typically refer to other items that share some features with it. For example, in spoken word recognition studies,

target items with many phonological neighbors are slower to be processed than those with fewer phonological neighbors (e.g., P. A. Luce & Pisoni, 1998). Such slow processing can be explained by inhibition caused by highly activated competitors. Frequent phonological neighbors are highly activated because their phonological features partially match the perceptual input and because they are frequent. Since those neighbors inhibit the activation of the target, items with frequent phonological neighbors are slower to be identified in spoken word recognition. In other studies of lexical access such as word production, it is also the case that highly activated neighbors interfere with the processing of the target items (See Chen & Mirman, 2012, for review.).

There is one more piece of online evidence from the visual world paradigm for lateral inhibition. Dahan, Magnuson, Tanenhaus, and Hogan (2001) found that activation of a non-target item slows down the recognition of the target item. They constructed triplets of two monosyllabic real words (W1 and W2) and one non-word (N3) which only differed from each other by the place of articulation of the final stop consonants (e.g. *net*, *neck*, and **nep*). One of the real words (W1) was the target item. The triplets were cross-spliced so that the initial portions before the final consonants were from each word in the triplet, but the final consonants were always from the target item. For example, in one triplet, stimuli in different conditions were all *net*, but the first portions [nɛ] were taken from *net* (W1W1), *neck* (W2W1), or *nep* (N3W1). Therefore, there were misleading coarticulatory cues in W2W1 and N3W1. Participants were presented with four pictures including the target item while they listened to W1W1, W2W1 or N3W1, and their eye movements were recorded. They found that participants fixated on the target item earlier in the N3W1 condition than in the W2W1 condition. That is, the target item was recognized more slowly when the first

portion of the stimuli was the same as a real word but not when it is from a non-word. This is consistent with models with lateral inhibition because the W2W1 stimuli temporarily activate W2, which can inhibit the activation of W1, while the N3W1 stimuli does not have a corresponding lexical representation to be activated. In fact, they ran a simulation using the TRACE model, which has lateral inhibition as one of its core components, and found that the model successfully predicts that W2W1 is activated more slowly than N3W1.

Thus, while the current study provides evidence for lateral inhibition in context-driven, top-down activation in the cloze task, there is existing evidence for lateral inhibition in bottom-up lexical activation. This is consistent with the idea that bottom-up and top-down activation (pre-activation) of lexical representations are caused by essentially the same process, but that only the sources of activation are different. This is a significant advance, given that there have been very few studies comparing the cloze task and bottom-up lexical activation.

Thus, this chapter has examined the linking hypothesis of the (speeded) cloze task so far. Based on these findings, the next section will examine the relationship between observable measures and the parameters of the model, which presumably reflects the underlying cognitive states or representations.

2.5 Linking cloze measures and the pre-activation mechanism

Building on the simulations conducted in this chapter, I explore more concrete relationships between the cloze measures and the underlying pre-activation in this section, through detailed analyses of the computational model and simulated data. The goal of this section is to demonstrate that different cloze measures reveal different aspects of pre-activation.

The nature of the Race model and the structures of the simulations in Chapter 2 enable us to consider the relationship between two key distinctions about pre-activation that prior studies did not make clearly. First, there is a distinction between strengths and activation levels of lexical candidates. Under the Race model, each lexical candidate has its own rate of pre-activation and accumulates pre-activation based on it. Some candidates such as *thief* in (8) are typically pre-activated quickly compared to less likely ones such as *smuggler* in (8). In the speeded cloze task, this is reflected in quicker and more frequent production of *thief*. Thus, there is a notion of how good each candidate usually is. I refer to typically quickly activated candidates such as *thief* in (8) as "strong" candidates, while typically slowly activated candidates such as *smuggler* in the same context as "weak" candidates. Whether a lexical candidate is strong is determined by its relationship with the context and the knowledge of the comprehender.

(8) The cop arrested the ...

The strengths of candidates are not identical to the transient activation levels of lexical candidates. As discussed in the previous section, it is not always the case that generally quickly activated candidates are activated quickly in every trial. Furthermore, even within a single trial, the same lexical item has different levels of activation at different time points of the pre-activation process as the activation accumulates. Thus, the amount of accumulated pre-activation of a lexical candidate can be considered as the transient state of the candidate as opposed to its stable profiles (strength).

I will argue in this section that strength is better reflected in the cloze RT. This is because RTs of a lexical candidate are affected by other competitors to a smaller degree compared to

the lexical candidate's cloze probability, and hence can better represent its properties. On the other hand, the (average) level of activation is better reflected in the cloze probability, because candidates are strongly activated in the cases they are winning races, but are weakly activated when they are not, due to lateral inhibition. This makes levels of activation resemble the cloze probabilities when they are averaged across winning and non-winning races.

There is another important distinction regarding the properties of contexts, often referred to as contextual constraints. While some key cloze measures reveal the strength or the activation levels of individual lexical candidates as described above (hence "candidate-level measures"), cloze data can also be used as measures of contexts, regarding the properties of sets of candidates that are generated given the context (hence "context-level measures"). One commonly used way to characterize contexts is the contextual or sentential constraints (e.g. Kutas & Hillyard, 1984; Federmeier & Kutas, 1999; Federmeier, Wlotko, De Ochoa-Dewald, & Kutas, 2007; Chow et al., 2018). High-constraint contexts have one or a very small number of highly predictable continuations, whereas low-constraint contexts do not have highly predictable continuations. This has been often measured by the highest offline cloze probability for the context (modal cloze probability), or by Shannon's entropy of the responses. However, the prior literature contrasting high and low constraint contexts has often conflated two different notions of whether there are one or more strong candidates available given the context (availability) and whether there is a dominant candidate (dominance).

I demonstrate that different properties of contexts are better reflected in different context-level cloze measures. The availability of candidates in the context is reflected in mean RTs of all responses given the context, whereas the dominance of one candidate in the context can be measured by the entropy of the cloze responses or the modal cloze probabilities.

The structures of the model in Simulation 2 enable the investigation of these key distinctions. The original simulation by Staub et al. (2015) only modeled the finishing times, which are the times for each item to reach the threshold for subsequent processes. This was based on the assumption that there is no interaction between the competing lexical candidates and hence the levels of activation were always proportional to the rates of activation. Given this assumption, it is difficult to make the distinction between the strengths and activation levels. On the other hand, my computational model implemented step-by-step accumulation of activation in order to model lateral inhibition. This naturally allows the rate of activation and the accumulated activation to be implemented and measured separately, and the structure is useful in examining the relationship between strengths/activation levels and cloze measures. Furthermore, the new computational simulations allow the investigation of context-level measures. Each of Staub et al. (2015)'s simulations was modeled using only a single set of candidates, which is comparable to a simulation of a speeded cloze task with a single context. On the other hand, the computational simulations in the previous section generated a number of sets of candidates, just like a simulated speeded cloze study with a number of contexts. This allows the comparison of different contexts which have different sets of candidates and the examination of the relationship between the properties of contexts and the context-level measures. Utilizing these features of the model, I analyze the model parameters and the simulated data in the rest of this chapter.

2.5.1 Candidate-level measures: strength vs activation level

I used the same model with the same parameters as Simulation 2 in this chapter to investigate the relationship between strengths and activation levels of lexical candidates and cloze mea-

sures. I generated 1000 sets of lexical candidates and ran 40 races for each set of candidates. As described in Chapter 2, each candidate obtained new activation in each cycle, and then it inhibited all its competitors depending on the amount of its total activation. The amount of pre-activation each candidate accumulates in each cycle was randomly determined based on normal distributions, where each candidate has a different μ parameter (i.e. the mean of the normal distribution). The μ value for each candidate was determined by randomly sampling from another normal distribution. The first lexical candidate whose activation exceeded the threshold of 1 was chosen as the cloze response. Cloze probabilities were obtained by aggregating those cloze responses and calculating the probability of each candidate winning the race. Cloze RTs were determined by the cycle when the winning candidate reached the threshold, assuming that each cycle is 25ms long and that it takes an additional 350ms after the completion of the race to initiate articulation (Indefrey & Levelt, 2004).

I focus on strengths and pre-activation levels as two important aspects of the pre-activation process. The strengths are implemented as the μ parameter for each lexical candidate, which determines the (average) rate of pre-activation. Candidates with greater μ values can be understood as strong lexical candidates in the context, which tend to accumulate activation more quickly compared to those with smaller μ values. I also kept track of the amount of activation for each candidate after each cycle. I took the activation level of each lexical candidate at the 16th and 32nd cycles, which correspond to 400ms and 800ms after the beginning of the race, and then averaged them across trials. The specific times of 400ms and 800ms were chosen rather arbitrarily, but activation levels at these points can be considered as states of each lexical candidate's pre-activation at different points in the race processes.

The relationship between the strengths, the activation levels, and the two candidate-level

cloze measures (the cloze RTs and the cloze probabilities) is plotted in Figure 16. Table 4 shows the Pearson correlation coefficients. As can be seen in the figure, strengths correlate with cloze RTs (the top left figure) better than the cloze probabilities (the bottom left figure). On the other hand, activation levels at both 400ms and 800ms after the race onset were better captured by the cloze probabilities (the bottom middle and bottom right figure).

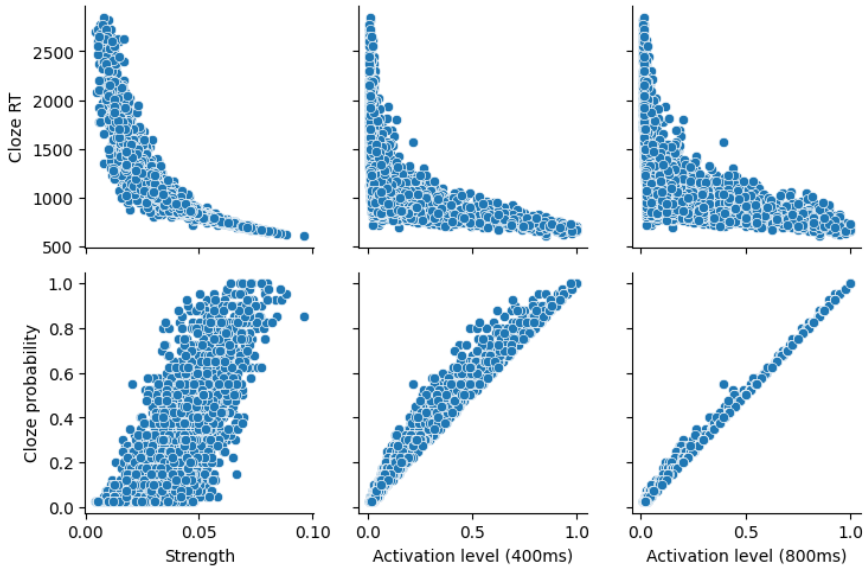


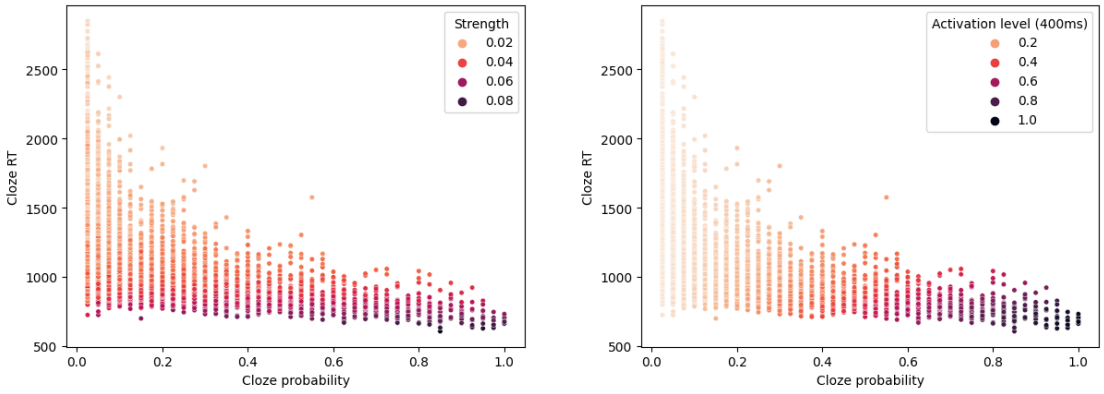
Figure 16: Illustrations of the relationship between the strength and activation levels of each candidate in the model (x-axes) and the simulated cloze measures of each candidate (y-axes). Each dot represents each lexical candidate. Cloze RTs are averaged for each candidate in the plot.

	Strength	Activation level (400ms)	Activation level (800ms)
Cloze RT	-0.82	-0.54	-0.55
Cloze probability	0.76	0.98	1.0

Table 4: The Pearson correlation coefficients of each candidate-level measure and the strength/activation level.

While there are also strong correlations between cloze probabilities and strengths (lower left figure), or between cloze RTs and activation levels (upper middle and right figures), further examination shows that these are mostly caused by correlation between the cloze RTs and cloze probabilities. Cloze probabilities and cloze RTs are plotted in Figures 17,

where Figure 17a shows the strength and Figure 17b shows the activation level as the color of the dots. As can be seen from the figure, the cloze RTs and probabilities correlate with each other. In Figure 17a, there is a clear pattern that the strength changes along the cloze RT (y-axis), while there is no clear effect along the cloze probability (x-axis). Conversely, in Figure 17b, the change in activation level is much clearer along the cloze probability compared to the cloze RT. Thus the apparent correlation between the cloze probability and the strength or between the cloze RT and the activation level seems to be mainly arising from the correlation between the cloze RT and cloze probability.



(a) Cloze measures and strengths **(b)** Cloze measures and activation levels

Figure 17: The relationship between the cloze RT-probability correlation and (a) the strengths and (b) the activation levels. The cloze measures are on the axes, and the strengths and the activation levels are shown by the colors of the dots.

Thus, the cloze RT seems to more closely correlate with the strength of the lexical candidate, while the cloze probability seems to be a better reflection of the activation levels. Lateral inhibition seems to contribute to the correlation between the cloze probability and the activation level. An individual race typically has one highly activated candidate and other very weakly activated candidates. This is because the highly activated candidate strongly inhibits the competitors and the inhibited candidates cannot inhibit their competitors strongly due

to their weak activation. Therefore, as time proceeds, a candidate that happened to accumulate a lot of activation at earlier times accumulates more and more activation and inhibits its competitors, whereas weakly activated candidates cannot accumulate much activation due to the inhibition from their competitors. Then the amount of activation averaged across different races is mainly determined by how often the candidate was the strongly activated one. In fact, the correlation between the cloze probability and the activation level becomes stronger at 800ms after the race onset compared to 400ms after the onset, presumably reflecting the stronger effects of lateral inhibition with more time.

On the other hand, the cloze RTs are less affected by lateral inhibition and they more transparently reflect the strength. This is because the cloze RTs are only measured when the lexical candidates win a race. In such cases, the winning lexical candidates receive relatively small amounts of inhibition from their competitors because the competitors are less activated than the winner. Hence the rates of activation are closely tied with the RTs, which are the times for each lexical candidate to win a race. Weak candidates are still more susceptible to inhibition than strong candidates, which can be seen in the weaker correlation between the strength and the cloze RT on the left side of the top left figure of Figure 16.

Thus, the examination of the simulated data shows that cloze RT is more sensitive to the strength of the lexical candidate, while cloze probability is more sensitive to the activation level and that these relationships are highly affected by lateral inhibition.

2.5.2 Context-level measures: Availability vs Dominance

In this subsection, I turn to properties of contexts and context-level measures. Contextual constraints are commonly used to characterize contexts. Contextual constraints refer to how

constraining the predictions generated given the context are, and high-constraint contexts have one very strongly predicted candidate. Contextual constraint is often operationalized using the highest cloze probability among the candidates, also known as the modal cloze probability, but sometimes Shannon's entropy of the cloze responses is used instead. Contextual constraints have been utilized for studies of how predictions are generated (e.g. Kutas & Hillyard, 1984; Federmeier & Kutas, 1999; Federmeier, 2007) or how failed predictions affect language processing (Van Petten & Luka, 2012).

Prior studies have often referred to two distinct notions with the same term of contextual constraint. One notion corresponds to the availability of candidates, or whether there is one or more strong candidates that typically accumulate activation quickly. The other corresponds to the dominance of one candidate, or how dominant the best candidate is compared to other competitors. These two notions have been conflated probably because prior studies have often assumed that subjective probabilistic distributions are directly represented in the human mind and are transparently reflected in lexical prediction measures. This has been a particularly common view for the cloze task, where researchers have often (implicitly) assumed that cloze responses are samples from the subjective probabilistic distributions of lexical candidates, and hence that cloze probabilities are relatively transparent reflections of those distributions. Under this view, a highly likely candidate has to be a dominant candidate as well. For example, if one candidate is 90% likely in the subjective probabilistic distributions, the other candidates must be less than 10% likely.

However, as discussed earlier in this chapter, this view is not supported for the (speeded) cloze task regarding the cloze RT data. Staub et al. (2015) showed that the sampling view cannot explain the two qualitative patterns of the human speeded cloze data: the negative

correlation between cloze RTs and cloze probabilities, and the shorter cloze RTs for candidates with faster competitors. Therefore, the subjective probabilistic distributions don't seem to be explicitly represented or to directly affect the race processes.

Unlike the sampling view, the two concepts of availability and dominance are not identical in the Race model, which is more consistent with the empirical data than the sampling view. Figure 18 shows hypothetical strengths of sets of lexical candidates, and each set is generated by a single context. In the bottom left panel of the figure, there are multiple candidates that have high rates of activation, which means that they are highly predicted. However, there is no dominant candidate because all of the candidates are highly likely. On the other hand, in the top right figure, there is a dominant candidate but there is no highly predicted candidate. Thus, it is important to make the distinction between the availability and the dominance when we refer to contextual constraints. Furthermore, prior studies have used the modal cloze probability as a measure of contextual constraints, but it was not clear whether it reflects the availability or the dominance.

I examined the relationship between the availability and the dominance of the lexical candidates and the context-level cloze measures. I used the highest value of μ (strength) among the set of lexical candidates to represent the availability (e.g. the left panels of Figure 18). I will henceforth refer to this as the maximum strength. If the context has one or more strong candidates, it has a large maximum strength, but if there is no strong candidate, the maximum strength should be small. On the other hand, the dominance was represented by the difference between the highest and the second-highest μ-value (e.g. the right panels of Figure 18). I will refer to this as the strength difference. If the context has one outstanding candidate, the context has a large strength difference. If the context does not have a single

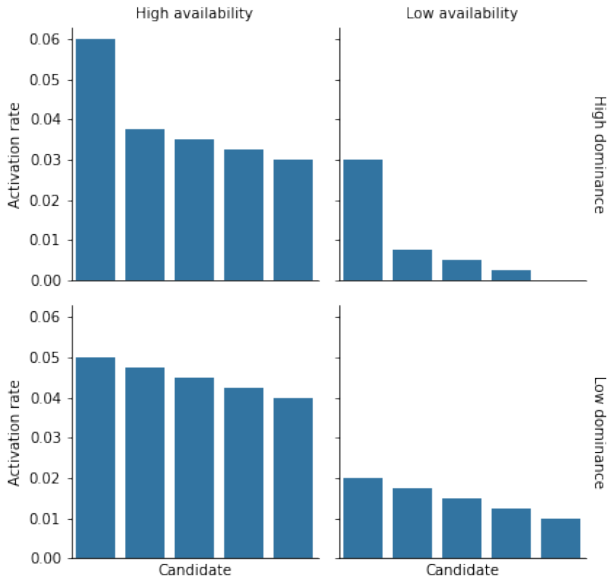


Figure 18: Distinction between the availability of candidates and the dominance of a candidate

outstanding candidate, the strength difference should be small. I used the modal cloze probability and the entropy of cloze responses as the context-level measures of availability or dominance as were used in the prior studies. Additionally, I also used the average cloze RT as a potential measure of availability, which is possible to obtain in the speeded cloze task and is likely to be revealing about the availability or the dominance.

Figure 19 compares the modal cloze probabilities, the mean RTs, and the entropy with the maximum strengths and the strength differences. Table 5 shows the Pearson correlation coefficients of the same combinations of the measures and the underlying availability and dominance. As the figures show, mean cloze RTs better represent the availability while the modal cloze probabilities or the entropy better represent the dominance.

There are some correlations between the other combinations, the mean RT and the dominance, or the modal cloze probability or entropy with the availability. This is because the strengths were generated based on a normal distribution, where high values were rare, and

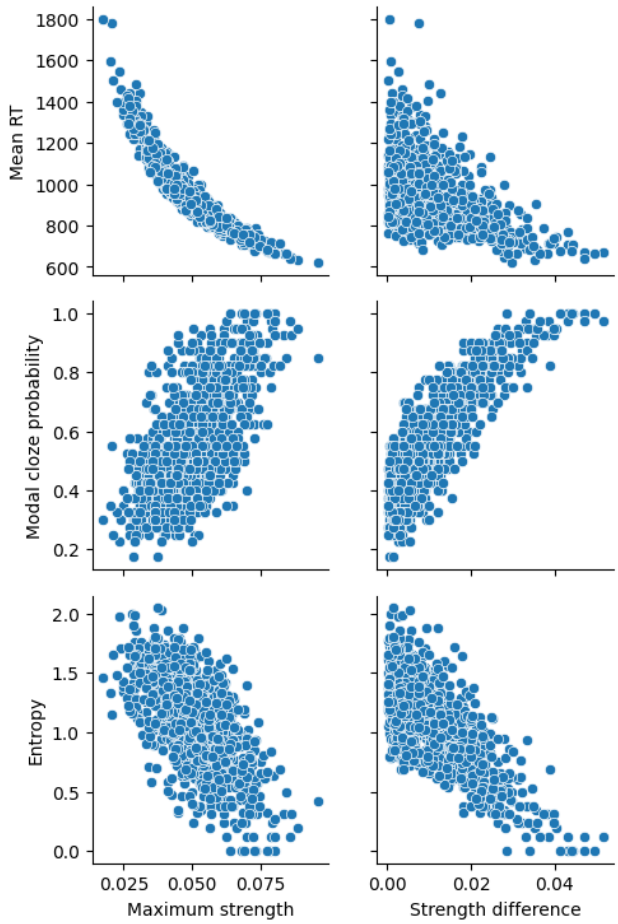
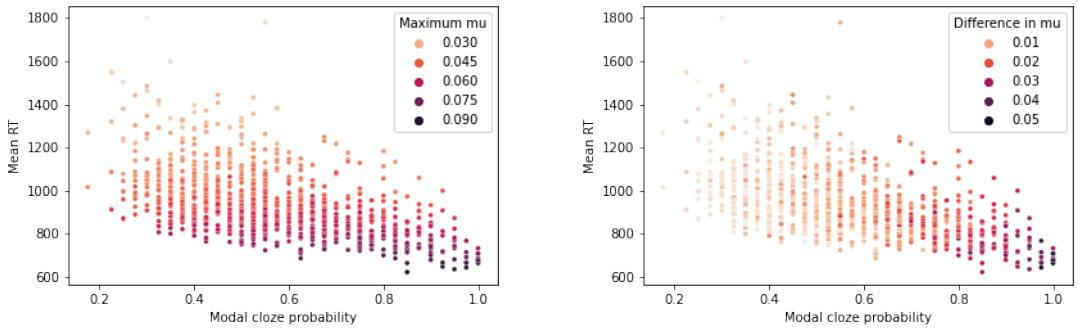


Figure 19: Illustrations of the relationship between the maximum strength and the strength difference of each context in the simulation (x-axes) and the simulated context-level cloze measures (mean RT, modal cloze probability, and entropy) (y-axes). Each dot represents each context.

	Maximum strength	Strength difference
Mean cloze RT	-0.92	-0.49
Modal cloze probability	0.63	0.85
Entropy	-0.64	-0.75

Table 5: The Pearson correlation coefficients of each context-level measure and the properties of sets of candidates.

hence sets of candidates with multiple high μ values were rare. This was done to follow human data where the modal cloze probability and the mean cloze RT values show a correlation. However, just like the previous subsection, these mostly arise from the correlations between the different context-level cloze measures. Figure 20b plots the mean cloze RT and the modal cloze probability on the axes, whereas the maximum strengths or the strength differences are plotted with the colors. The maximum strengths change with the mean cloze RTs but not with the modal cloze probability. On the other hand, the strength differences change with the modal cloze probabilities but not with the mean cloze RTs.



(a) Context-level cloze measures and availability **(b)** Context-level cloze measures and dominance

Figure 20: The relationship between the correlation of context-level cloze measures (mean RT, modal cloze probability, and entropy) and (a) the maximum strengths and (b) the strength differences. The cloze measures are on the axes, and the maximum strengths and the strength differences are shown by the colors of the dots.

2.5.3 Discussion

The simulations showed that different cloze measures are more closely related to different aspects of pre-activation. For the candidate-level measures, cloze RTs correlate with the strength, while cloze probabilities correlate with the amount of accumulated activation. For the context-level measures, mean cloze RTs reflect the availability of candidates, or the

strength of the best candidate. On the other hand, modal cloze probabilities or entropy of cloze responses better reflect the dominance of a candidate.

Cloze measures, especially the cloze probabilities and the modal cloze probabilities, have been frequently used in prior literature, but the current study provides new insights about what different cloze measures reflect and how they can be utilized. First, cloze RTs of a lexical candidate can be used as a unique window into the strength of the lexical candidate. Cloze RTs of a lexical candidate only reflect the cases where it was highly activated and won a race. The candidate's activation in those cases is less susceptible to inhibition from its competitors than in the cases where they were not highly activated. Therefore, cloze RTs better reflect the strength of the lexical candidate, which is difficult to measure using other measures of pre-activation. The strength reflects the stable profiles of the lexical candidate (i.e. how fast it is typically activated given the context) that is determined by contextual information and stored knowledge in human language processing. Therefore, a measure of strength is potentially informative in the investigation of how the rate of pre-activation of each lexical candidate is determined by contextual information and stored knowledge.

Second, the current study highlights the distinction between availability and dominance. They have often been referred to as contextual constraints without clear distinctions. This might make researchers choose the inappropriate measure, such as using modal cloze probabilities to measure the availability. Such inappropriate use of context-level cloze measures have limited influence in some studies because the modal cloze probabilities and the mean cloze RTs correlate with each other (Figure 21: Staub's Experiment 2). However, there are some new studies showing that the modal cloze probabilities and the mean cloze RTs do not consistently show strong correlations across developmental stages. In Lee et al. (2023),

we conducted a speeded cloze study with 4-12-year-old children and found that their modal cloze probabilities and mean cloze RTs do not align with each other as closely as in adults' data. This suggests that children's knowledge or pre-activation mechanisms have different features from adult's knowledge or pre-activation mechanisms that allow contexts with high dominance and low availability, or with low dominance with high availability.

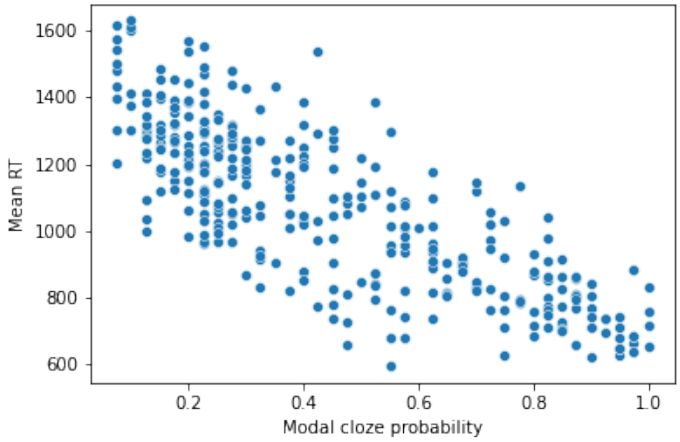


Figure 21: The correlation between the modal cloze probability and the mean cloze RT in Staub et al.'s Experiment 2.

In this chapter, I examined the Race model, the linking hypothesis of the speeded cloze task proposed by Staub et al. (2015). Although it successfully explains some important qualitative patterns in human speeded cloze data, I demonstrated that they make inaccurate predictions about quantitative patterns in the same data. I showed that an additional but reasonable assumption of lateral inhibition in the pre-activation processes can explain the quantitative pattern, preserving the race processes. Building on the new insights regarding lateral inhibition and the updated model, I further examined how the strength of lexical candidates and the activation levels of candidates are reflected in candidate-level cloze measures (cloze RT and cloze probability), as well as how profiles of contexts (availability

and dominance) are related to the context-level cloze measures (mean cloze RT, modal cloze probability, and cloze entropy). This serves as a basis for the subsequent study addressing a mismatch between cloze probabilities and N400 amplitudes in argument role reversals (Chapter 3), the study using context-wise variability of candidates' strengths measured by the speeded cloze task (Chapter 4), and an investigation of contrasts between stimulus sets using the speeded cloze task (Chapter 6).

3 Argument roles reversals in EEG vs Cloze

The previous chapter examined the linking hypothesis of the (speeded) cloze task. I addressed an issue for the original Race model proposed by Staub et al. (2015) by keeping the linking hypothesis but assuming lateral inhibition in pre-activation. I further examined the relationship between different cloze measures and the underlying pre-activation mechanisms and showed that different cloze measures reflect pre-activation in different ways.

Under the Simple Activation Model, cloze measures should align with other measures of pre-activation (Figure 22), because different pre-activation measures reflect the shared pre-activation, and because there is a relatively transparent relationship between pre-activation and the measures as introduced in Chapter 1. Specifically, the relationship between pre-activation and cloze meausures were further examined in Chapter 2. Given these assumptions and the study in Chapter 2, when there is a large contrast in cloze probabilities, other measures should also show clear contrasts. In fact, such alignment of cloze measures and other pre-activation measures has been found repeatedly (N400: e.g., Kutas & Hillyard, 1984; Federmeier, 2007; Wlotko & Federmeier, 2012; Nieuwland et al., 2020; eye-tracking while reading: e.g., Ehrlich & Rayner, 1981; Staub, 2015; lexical decision task: e.g., Kleiman, 1980; read aloud task: e.g., Seidenberg et al., 1984). While these studies compared untimed cloze probabilities with other pre-activation measures, the relationships are likely to hold with the speeded cloze probabilities as well because untimed and speeded cloze probabilities show extremely high correlation with each other (Staub et al., 2015).

However, there are some notable cases where the pre-activation measures may not align with each other. One of them is argument role reversals, such as (9). *Serve* should be

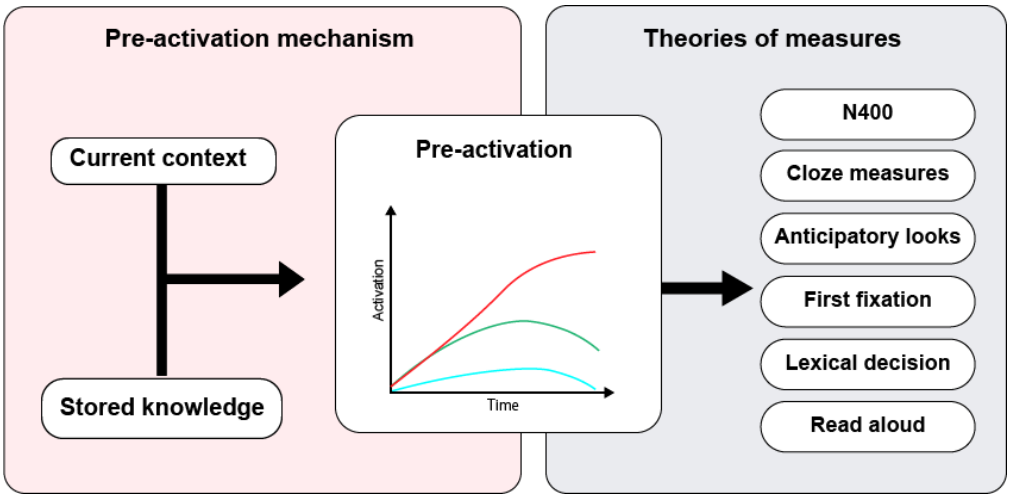


Figure 22: The Simple Activation Model of pre-activation, where the shared underlying pre-activation .

predicted in (9a) but not in (9b) given the roles of the arguments *waitress* and *customer*. Unsurprisingly, it has been robustly observed that untimed cloze probabilities show a clear contrast between the two *serve* (e.g., Chow, Smith, et al., 2016). However, a number of studies have shown that N400 amplitudes are matched for the two instances of *serve* despite the apparent contrast in predictability (e.g., Kolk et al., 2003; Chow, Smith, et al., 2016), with some converging evidence from other measures (anticipatory looks Kukona et al., 2011; first fixation times: Burnsky, 2022). This has been taken to show that argument roles do not affect pre-activation.

(9) a. The restaurant owner forgot which customer the waitress had served.

b. The restaurant owner forgot which waitress the customer had served.

Such cases of mismatch between pre-activation measures are useful in my attempt to understand lexical pre-activation by updating the Simple Activation Model. First, it informs us of what information can affect pre-activation generation in what way. Since the different

measures tell different stories about whether argument roles can impact pre-activation, we may be able to identify the roles of argument roles in the pre-activation mechanism by reconciling the apparently conflicting findings. Second, the mismatch calls into question the model's assumptions about theories of measures and it provides clues to refining the model. The mismatch between measures cannot be explained by the assumption of the Simple Activation Model that different measures transparently reflect the same shared pre-activation. Therefore, the linking hypotheses needs to be revised by identifying the sources of mismatch between measures.

Some researchers have proposed that the discrepancy between pre-activation measures arises from differences in timing (Chow et al., 2018). In most prior studies, online pre-activation measures such as N400 amplitudes reflecting real-time language processing are not sensitive to argument roles, while offline measures given ample time and processing resources, such as the untimed cloze task, show sensitivity to argument roles. However, there is some suggestive evidence from a speeded cloze study that the timing difference may not explain the contrast between the different pre-activation measures (Chow et al., 2015). I build on this preliminary evidence in the current chapter.

In this chapter, I directly compare a Japanese EEG experiment and a speeded cloze study to examine the discrepancy between measures and the effect of timing manipulation[2]. I reanalyze Japanese EEG data collected by Momma (2016) and run a speeded cloze experiment to investigate the contributions of task and timing differences to argument role sensitivity in pre-activation. Using maximally simple stimuli such as (10), the EEG study compared N400 amplitudes at different moments by controlling the timing of the presentation of the target

[2]This study was conducted as a collaboration with Colin Phillips and Shota Momma.

verbs. The cloze study used the same set of stimuli and tested the contrast between the two pre-activation measures by inducing responses within specific time windows that matched the EEG study. I show that the difference between the cloze studies and N400 studies remains, even after matching the timing profiles of experiments.

(10) a. Hati-ga sas-u.
Bee-NOM sting-PRES
'A/the bee stings something.'

b. Hati-o sas-u.
Bee-ACC sting-PRES.
'Somethings stings a/the bee.'

3.1 Prior studies on argument role reversals

Many EEG studies have suggested that predictions are insensitive to argument roles. Studies in different languages reported that the N400 component of ERP, which is considered to reflect the amount of pre-activation of the presented lexical item, is matched in amplitude for continuations such as *served* in (9), despite the clear contrast in predictability (English: Kuperberg et al., 2003; Kim & Osterhout, 2005; Chow, Smith, et al., 2016; Dutch: Kolk et al., 2003; Hoeks et al., 2004; Mandarin: Chow et al., 2018; Liao et al., 2022; Japanese: Oishi & Tsutomu, 2010; However, see Bornkessel-Schlesewsky et al., 2011 for exceptions). There are also some consistent findings from eye-tracking studies (Visual world paradigm: Kukona et al., 2011; eye-tracking while reading: Burnsky, 2022). Given the linking assumptions for N400 amplitudes, the previous findings showing the absence of N400 contrasts between role-appropriate and inappropriate continuations such as *served* in (9a) and (9b) have been taken to show that people's predictions about upcoming verbs are unconstrained by the argument

roles of items in the context. (e.g., Chow, Momma, et al., 2016; Kuperberg, 2016; Brouwer et al., 2017).

It is important to note that there are other interpretations of the absence of N400 effects in argument role reversals. Some researchers argue that it arises from heuristic semantic integration that ignores argument roles, rather than from lexical prediction, assuming that N400 amplitudes reflect the effort involved in integrating retrieved lexical representations into utterance meaning representations (e.g., Kuperberg et al., 2003; van Herten, Kolk, & Chwilla, 2005) or the detection of semantic anomaly (e.g., Kim & Osterhout, 2005; Li & Ettinger, 2023). In this account, (9b) is interpreted as (9a) at least temporarily, and hence there is no semantic anomaly. Therefore, the cost for integration of *serve* is matched for the two sentences. It is unlikely that comprehenders retain such non-literal interpretations, because participants accurately judge the argument role reversal sentences to be unacceptable in 80-95% of trials in most of the EEG studies of argument role reversals cited above. Therefore, it must be the case that any non-literal heuristic interpretations of argument role reversals are transient and are overridden by literal interpretations eventually (e.g., van Herten et al., 2005). I will address this alternative account in Section 3.5.

While many studies, as listed above, have repeatedly found insensitivity to argument role reversals in N400 amplitudes, certain task manipulations seem to make lexical prediction sensitive to argument roles. It is known that cloze probability is sensitive to argument roles, and verbs have a higher cloze probability in the contexts where they are appropriate regarding the roles (e.g. *served* in (9a)) than when they are inappropriate (e.g. *served* in (9b)) (e.g., Chow, Smith, et al., 2016).

However, the contrast between pre-activation measures might arise from differences in

timing rather than from the tasks themselves. Because most cloze studies are untimed, researchers have often assumed that cloze probabilities do not reflect online sentence processing, but have rather taken them as reflecting the upper bound of people's ability to use contextual information, given a lot of time and cognitive resources. This led them to assume that lexical pre-activation is differently sensitive to argument roles depending on the timing. In an early stage, pre-activation is driven by a role-insensitive mechanism, which is reflected in the N400 studies. On the other hand, in a later stage, pre-activation is driven by a role-sensitive mechanism which is reflected in the offline cloze studies.

Consistent with this view, Chow et al. (2018) argued that people activate upcoming verbs in a role-sensitive fashion if additional time is provided for prediction. They contrasted Mandarin SOV sentence pairs such as (11) and (12) including argument role reversals (11b and 12b). They found matched N400 amplitudes for the verb *zhuale* 'arrest' in (11a) and (11b), consistent with other studies. However, they found a larger N400 for the same verb in (12a) than in (12b) when prepositional phrases (PPs) were moved from the beginning of the sentence to the position between the second argument and the verb. While there are some concerns that (12) may sound somewhat unnatural to native Mandarin speakers, the study suggests that insensitivity to argument roles may be conditioned on timing, and that the role-sensitivity of prediction might change dynamically after the presentation of the context arguments but before the presentation of the target verb. There is also a converging finding from Liao et al. (2022).

(11) a. Zai shangxingqi jingcha ba xiaotou zhua-le
ZAI last.week cop BA thief arrest...
'Last week the cop arrested the thief...'

b. Zai shangxingqi xiaotou ba jingcha zhua-le
ZAI last.week thief BA cop arrest...
'Last week the thief arrested the cop...'

(12) a. Jingcha ba xiaotou zai shangxingqi zhua-le
Cop BA thief ZAI last.week arrested...
'Last week the cop arrested the thief...'

b. Xiaotou ba jingcha zai shangxingqi zhua-le
Thief BA cop ZAI last.week arrested...
'Last week the thief arrested the cop...'

However, an unpublished cloze study by Chow et al. (2015) suggests that the online-offline contrast might not explain the difference between the cloze task and online measures of context-driven activation. They conducted a cloze task, but imposed deadlines for production by asking participants to produce a continuation within certain time limits, in an attempt to elicit role-inappropriate responses in the cloze task. Results showed that people are far more likely to produce role-appropriate continuations than role-inappropriate continuations, even in a cloze task under time pressure. This suggests that cloze data show sensitivity to argument roles not because they are offline measures, and that people can show early sensitivity to argument roles, in at least some experimental tasks. However, since they were seeking evidence for the parallelism between the cloze and the EEG studies, they focused more on the small number of role-inappropriate responses and did not pay much attention to the relative scarcity of those role-inappropriate responses compared to role-appropriate responses. Furthermore, we can only draw limited conclusions from this study because Chow et al. did not closely match the language or timing with prior EEG studies that showed timing effects (Chow et al., 2018).

3.2 Current study

This study aims to address timing and task effects in argument role reversals. I reanalyzed the data from a Japanese EEG study on argument role reversals in Momma (2016) that manipulated the timing of verb presentation with minimal changes otherwise. I also conducted a speeded cloze study that was matched with the EEG study in terms of the stimuli and timing to compare different measures at the same timing. The characteristics of Japanese allowed us to create maximally simple stimuli such as (10), which were ideal for these purposes. The sentences consisted of one context noun with a case marker and a target verb. Since Japanese is a verb-final language that allows free dropping of arguments, both (10a) and (10b) are grammatical. The role-appropriate (10a) and inappropriate (10b) differ only in the case markers on the context nouns. Participants had relatively easy access to argument role information because the explicitly marked cases strongly correlate with argument roles. The agent role is typically assigned to arguments with the nominative case (NOM), whereas those with the accusative case (ACC) usually receive the patient role. Of course, case markers and argument roles are not perfectly correlated in Japanese. For example, passive subjects are marked with the nominative case. But there is a strong probabilistic association in Japanese, and there was a perfect correlation within the current study. Importantly, the simplicity of the materials allowed Momma (2016)'s EEG study to manipulate timing by simply varying the interval between the context noun and the target verb, without changing any other aspects of the stimuli. This makes a stronger case for the timing effects than Chow et al. (2018), where the differences in the positions of the PPs could, in principle, have resulted in slightly different predictions.

In the EEG study, participants were presented with role-appropriate and inappropriate stimuli such as (10). The stimulus onset asynchrony (SOA) between the context noun and the target verb was 800ms in the short condition and 1200ms in the long condition (Figure 23). Momma compared the N400 contrasts between the target verbs as role-appropriate continuations, such as *sting* in (10a), or as role-inappropriate continuations, such as *sting* in (10b), with different delays following the context nouns.

In the cloze study, I presented participants with the same contexts as the EEG study, but without the target verbs. Instead, participants were asked to produce a natural verb to continue each context noun within certain time windows that matched the EEG study. The time windows were before 1200ms following the context noun in the short condition and between 1600ms and 2800ms in the long condition. Given that it is estimated to take about 350ms to initiate articulation after completing lexical access (Indefrey & Levelt, 2004), I estimated that participants had up to 850ms for lexical activation in the short condition and 1250-2450ms in the long condition, which matches the EEG timing manipulation (Figure 23). This timing manipulation may not have perfectly matched the EEG and the cloze studies, because the SOA in the EEG study might have underestimated the time available for context-driven lexical activation. Specifically, participants might continue activating possible upcoming verbs based on the contextual information even after the verb presentation until the bottom-up processing of the verb is completed. However, this possibility has minimal influence on the interpretation of the data because this additional time should make N400 amplitude more sensitive to argument roles due to the additional time, if there is any effect. If I find that people are sensitive to argument roles in the cloze task but not in the EEG experiment, as expected, the concern about timing alignment would be moot. Thus, the

design allowed the comparison of people's pre-activation at different moments in the same EEG study, and at the same moment in different paradigms.

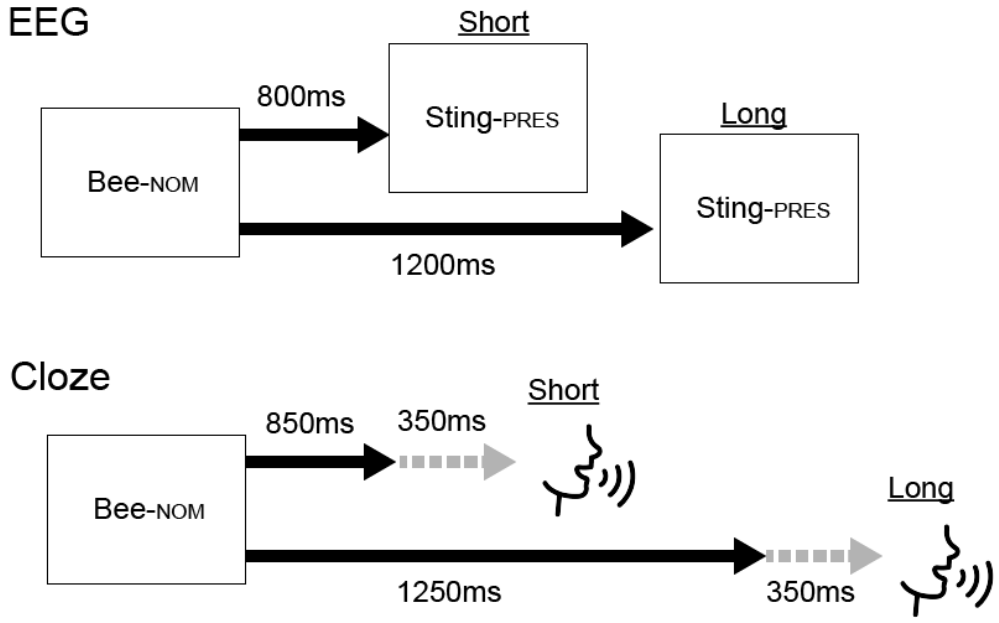


Figure 23: The timing manipulation in the EEG and speeded cloze studies. The SOA was 800ms in the short condition and 1200ms in the long condition in the EEG study. In the cloze task, if a continuation is produced 1200ms after the onset of the context (short condition) there should have been 850ms available for activation. If a continuation was produced 1600ms following the context, there should have been 1250ms for activation.

3.3 Experiment 3-1: EEG experiment

3.3.1 Methods

In this subsection, I describe how Momma (2016) collected the EEG data and how I analyzed the data.

Participants

27 participants were recruited at Hiroshima University. I excluded three participants accord-

ing to a data quality criterion described below, leaving 24 (12 males and 12 females; average age = 21.2). All participants were right-handed, and were native speakers of Japanese without any significant early exposure to other languages.

Materials

Momma (2016) created 160 plausible sentences with one case-marked context noun and one target verb. 80 of them had the nominative (NOM) case marker *-ga* such as (13a) and the other 80 had the accusative (ACC) case marker *-o* such as (14a). These sentences were the plausible condition in the experiment. Then, he created the implausible condition by switching the cases of those context nouns ((13b) and (14b)) in the plausible condition. All the context nouns were animate, and all the target verbs were transitive verbs that can have multiple arguments. He also manipulated the timing of the presentation of the target verb. In the short condition, the target verb was presented 800ms after the onset of the context noun. The interval was 1200ms in the long condition. These two manipulations resulted in four conditions in total: plausible-short, plausible-long, implausible-short, and implausible-long. The 160 experimental items in the four conditions were distributed across four lists using a Latin-Square design so that the same participant saw each experimental item just once. In addition to the experimental items, Momma also created 80 plausible and 80 implausible filler sentences, which were also presented with an SOA of either 800ms or 1200ms.

(13) a. Hati-ga sas-u
Bee-NOM sting-PRES
'A/the bee stings something.'

b. Hati-o sas-u
Bee-ACC sting-PRES

'Somethings stings a/the bee.'

(14) a. Hae-o tatak-u
Fly-ACC swat-PRES
'Something swats a/the fly.'

b. Hae-ga tatak-u
Fly-NOM swat-PRES
'A/the fly swats something.'

One item was incorrectly labeled in the stimuli. This resulted in some participants seeing one more or fewer plausible items than implausible items. This is unlikely to have affected the results because I relabeled these stimuli for the analyses, and it is unlikely that this small shift in the ratio of plausible and implausible stimuli would result in observable effects.

Procedure

The 320 sentences were divided into four blocks divided by short breaks. Each sentence was visually presented to participants word-by-word using PsychoPy (Peirce, 2007) while brain activity was recorded via EEG. In each trial, a fixation cross was presented at the center of the screen for 500ms, and a blank screen was presented for the next 400ms. The context noun and the target verb were presented word-by-word separated by a blank screen. Each word remained on the screen for 400ms. The interval of the blank screen was 400ms in the short condition and 800ms in the long condition. Therefore, combined with the time the context noun was on the screen, the SOA was 800ms in the short condition and 1200ms in the long condition. Participants were asked about the plausibility of each sentence 1000ms after the offset of the target verb. They provided answers by pressing keys.

Eletcrophysiological recording

32 sintered Ag/AgCl passive electrodes were placed on participants' scalps in an EEG recording cap (Brain Products GmbH Easy Cap 40 Asian cut), according to the 10-20 system. One of the 32 electrodes was placed below the right eye to monitor eye movements and blinking. One electrode was used as ground, and another was used as a reference electrode and was attached to the nose. Impedances were kept below 5kOhms for all scalp electrode sites. The EEG was sampled at 250 Hz.

Data analysis

The recorded EEG signals were preprocessed and averaged ERPs were obtained using EEGLAB (Delorme & Makeig, 2004) and ERPLAB (Lopez-Calderon & Luck, 2014) following (Luck, 2021)'s pipeline. The EEG signal was bandpass-filtered between 0.1 and 30Hz. Eye blinks and movements were removed by independent component analysis (ICA). Non-responding channels were excluded from this artifact correction procedure and were interpolated from neighboring channels using spherical splines after the correction. Then the EEG was segmented into 1000ms intervals, starting 200ms before the target verb onset and ending 800ms after the verb onset. A baseline correction was applied to the 200ms epoch prior to the target verb onset. Trials with EEG above 150 µV or below -150 µV were marked for rejection. Additionally, artifacts were detected and rejected using the moving window peak-to-peak algorithm with a window size of 200ms, a window step of 100ms, and a threshold of 125 µV. Three participants who had a 20% or greater rejection rate were excluded from further analyses. The rejection rate among the remaining participants was 1.1% on average.

I first independently compared the plausible and implausible conditions for each SOA

(short vs long). The EEG data was analyzed with nonparametric cluster-based permutation tests using the FieldTrip toolbox (Oostenveld, Fries, Maris, & Schoffelen, 2011). I obtained individual participant averages at each electrode for each time point from 300 to 800ms after the onset of the target verb, and computed t-values for each electrode and time. I then selected spatially and temporally adjacent samples based on the t-values to form clusters ($\alpha = 0.05$). The summed t-values within each cluster were used as the cluster-level test statistic. Then I randomly permuted the condition labels 1000 times and obtained the maximum of the cluster-level test statistics for each iteration. The test statistics of the data were compared against this randomly-generated distribution, and the p-value of each cluster is the proportion of the generated test statistics greater than that of the cluster. I used the α-level of 0.025 for two-tailed tests. I examined if there was a statistically significant negative cluster in the 350-450ms time window, which is commonly used for N400 analyses. In addition to the analyses within each SOA, I also analyzed the interaction between plausibility and SOA by comparing the difference between the plausible and implausible conditions for each SOA and performing the same analysis as above (i.e., the difference between implausible-long and plausible-long compared against the difference between implausible-short and implausible-short).

3.3.2 Results

Behavioral data

The overall accuracy of the plausibility judgment task was 93.4%, including both experimental and filler items. More specifically, the accuracy was 96.4% for the plausible-short condition, 90.5% for the implausible-short condition, 96.8% for the plausible-long condition, and 89.8% for the implausible-long condition (all averages including filler items). A bino-

mial generalized mixed-effects model with maximal random effects structures that converged revealed a significant effect of plausibility ($\beta = 1.3417, p < .001$), presumably reflecting a response bias towards plausible judgments. On the other hand, there was no statistically significant effect of SOA or interaction between plausibility and SOA.

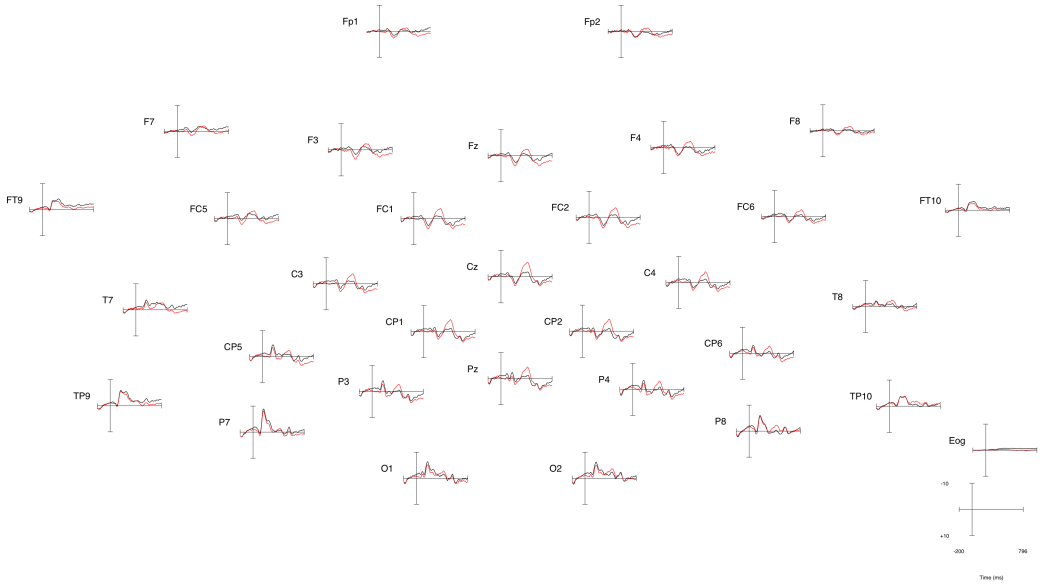
EEG data

Filler items

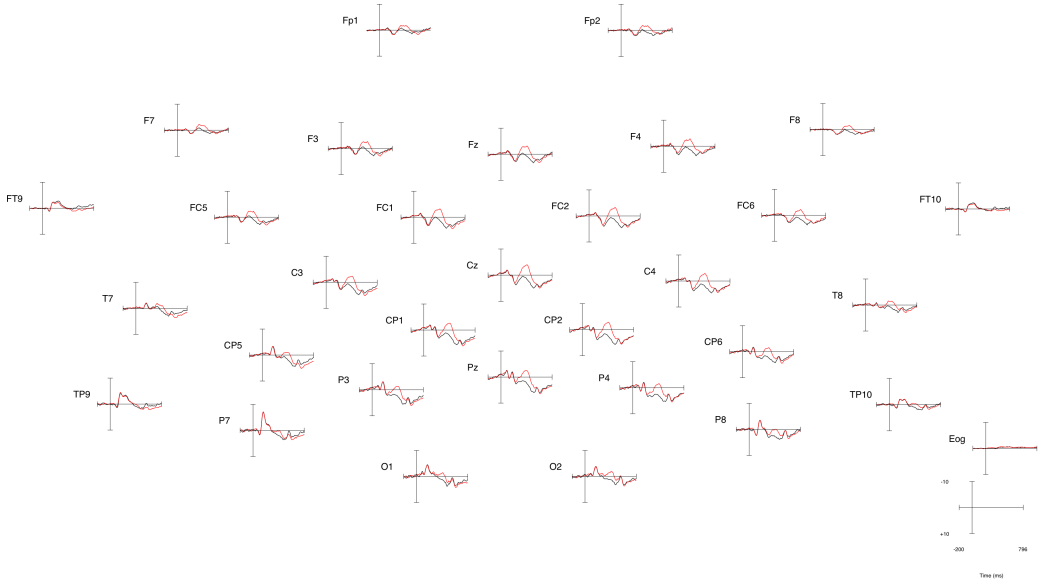
60 trials were excluded from the analyses because they were erroneously shown to participants multiple times. This left 3783 filler trials to be analyzed. Figure 24 shows the ERPs of the filler stimuli. As illustrated by the figures, classic N400 effects were observed in the comparison between the plausible and implausible conditions, both for the short and long conditions. Figure 25 shows topographic maps of ERPs at 300-750ms after the target verb onset, where each map is based on the average of a 50ms window. Significant clusters are indicated with stars. As illustrated in the figure, cluster-based permutation tests found significant negative clusters in the N400 time window (350-450ms) both in the short condition ($p = .010$) and the long condition ($p < .001$). That is, the implausible conditions were more negative than the plausible conditions in this cluster. A positive cluster was unexpectedly found in the late time window (570-800ms) in the short condition ($p = .016$). No positive or negative clusters were found for the interaction analysis.

Experimental items

Figure 26 shows the ERP waveforms for the experimental items. As Figure 27 illustrates, there was no statistically significant positive or negative cluster found in the short condition.



(a) Short condition (Filler)



(b) Long condition (Filler)

Figure 24: ERPs of (a) the short condition and (b) the long condition of the filler items. Negative voltages are plotted upwards and positive voltages are plotted downwards. This is the same for other ERP plots in this chapter.

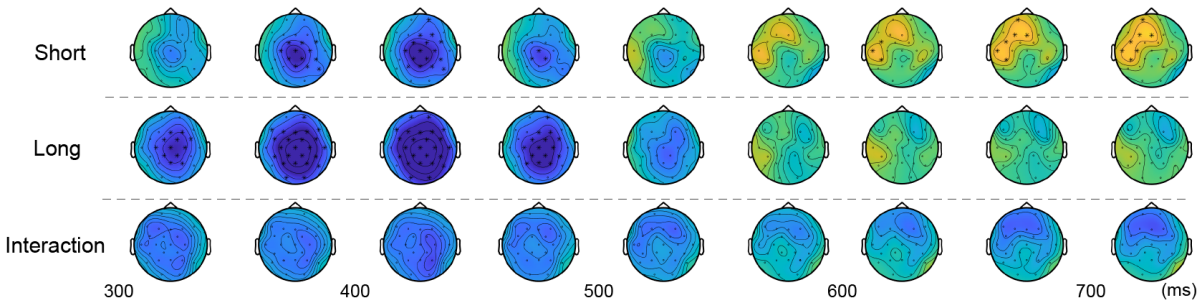


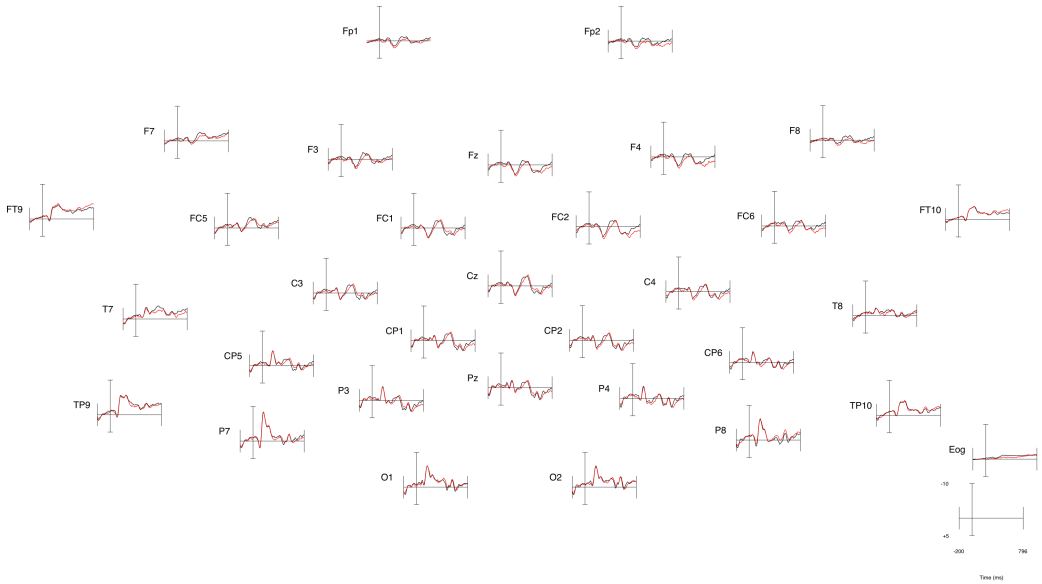
Figure 25: Topographic maps of ERPs of the filler items. The short condition was subtracted from the long condition for the interaction. Stars (*) indicate members of significant clusters. Blue indicates negative voltages and yellow indicates positive voltages. This format is consistent in other topographic plots of ERPs in this chapter.

On the other hand, there was a significant negative cluster in the N400 time window in the long condition ($p = .007$). Additionally, there was a significant negative cluster in the N400 time window in the interaction analysis ($p = .013$). These show that the long-implausible condition showed more negativity than the long-plausible condition in the N400 time window, but there was no evidence for such a contrast in the short conditions.

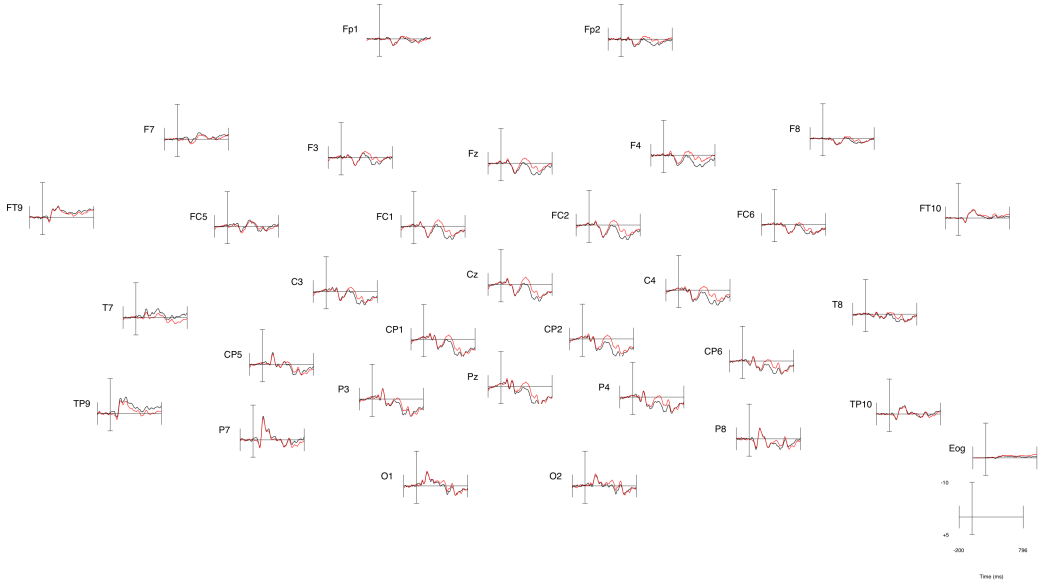
3.3.3 Discussion for the EEG experiment

The EEG study showed an N400 contrast between the plausible-long condition and the implausible-short condition, but not between the plausible-short condition and the implausible-short condition. The lack of an N400 effect in the short conditions is consistent with many previous EEG studies (e.g., Kuperberg et al., 2003; Kim & Osterhout, 2005; Chow, Smith, et al., 2016). The presence of an N400 effect in the long condition replicates Chow et al. (2018). Thus N400 amplitude at the verb is insensitive to argument roles in the context when the interval between the context and the target verb is short, but it shows sensitivity with an additional 400ms before verb presentation. Note that this additional time (400ms) is much shorter than the additional time added by moving a prepositional phrase (1200ms) in Chow

et al. (2018) and Liao et al. (2022).



(a) Short condition (Filler)



(b) Long condition (Filler)

Figure 26: ERPs of (a) the short condition and (b) the long condition of the filler items.

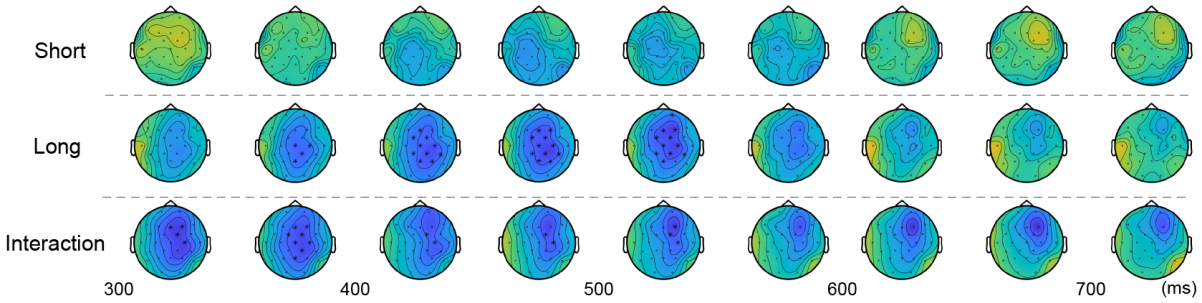


Figure 27: Topographic maps of ERPs of the experimental items. Stars (*) indicate members of significant clusters.

Nominative-plausible and accusative-plausible items

Reversal anomalies with nominative context nouns might, in principle, be able to be repaired by passives or causatives. For example (15b) is an argument role reversal sentence, but (15c) is a plausible sentence, where the nominative noun is the patient. Therefore the lack of N400 effects in the short condition might be mainly caused by sentence pairs such as (15a) and (15b), where the same verb could be used in a plausible sentence. On the other hand, it is impossible to turn the argument role reversal sentence (16b) into a plausible sentence through suffixation.

(15) a. Hae-o tatak-u
Fly-ACC swat-PRES
'Something swats a/the fly.'

b. Hae-ga tatak-u
Fly-NOM swat-PRES
'A/the fly swats something.'

c. Hae-ga tatak-arer-u
Fly-NOM swat-passive-PRES
'A/the fly is swatted by something.'

(16) a. Hati-ga sas-u
Bee-NOM sting-PRES

'A/the bee stings something.'

b. Hati-o sas-u
Bee-ACC sting-PRES
'Somethings stings a/the bee.'

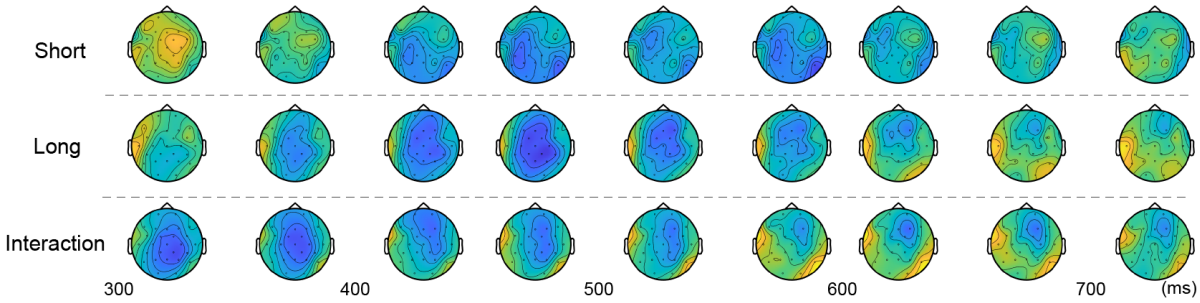
In order to address this concern, I also conducted the same analyses as above but separately for accusative-plausible stimuli such as (15) and nominative-plausible stimuli such as (16). There were no statistically significant positive or negative clusters for any analyses, presumably due to the reduced data points, but there was a visually similar pattern in this partial data analysis (Figure 28). I take this to be consistent with the idea that the absence of an N400 effect in the short condition and its presence in the long condition is not caused specifically by one of the two types of stimuli. This suggests that the possibility of sentences such as (15c) did not contribute significantly to the results.

3.4 Experiment 3-2: Speeded cloze experiment

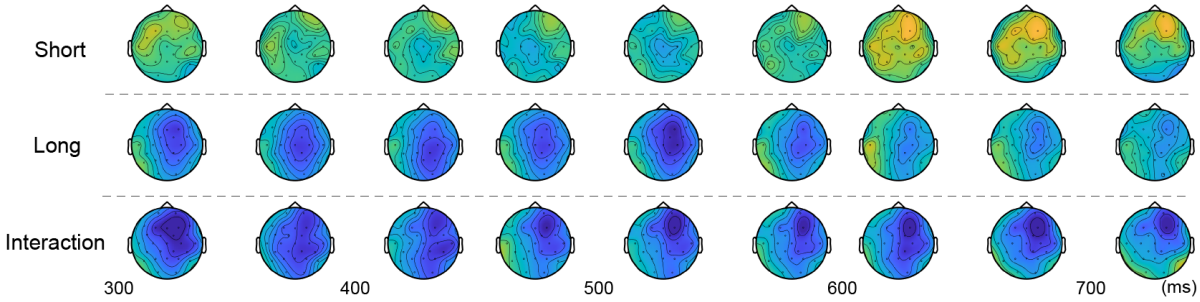
3.4.1 Methods

Participants

80 native Japanese speakers were recruited online on the crowdsourcing platform CrowdWorks. Participants received monetary compensation. Seven participants' data were not analyzed due to missing or unclear recordings, or not performing the task. Additionally, 19 participants were excluded from the data due to unsatisfactory performance as described in the Exclusion section. This left 54 participants for analysis. (39 females and 14 males, 1 did not provide gender information; average age = 41.1, 2 did not provide age information.)



(a) Accusative-plausible stimuli



(b) Nominative-plausible stimuli

Figure 28: Topographic maps of ERPs of the (a) accusative-plausible and (b) nominative-plausible stimuli. There were no significant clusters in either analysis, but the visual pattern in the nominative-plausible stimuli was similar to the accusative-plausible stimuli.

Materials

We used 159 case-marked context nouns from the 160 experimental item pairs in the EEG experiment. Unlike the EEG experiment, each context noun did not appear with a target verb, and participants instead provided a verb that was a natural continuation of the presented context. The nouns were presented with either nominative or accusative case. One stimulus was replaced with another item because the same context noun appeared twice in the EEG experiment. This context noun caused minimal problems for the EEG experiment because it appeared with two different verbs, but these two experimental items would have been identical in the cloze task where verbs are not presented. No fillers were used for this experiment because we did not directly present reversal anomalies in the cloze task and therefore there was minimal risk for participants to realize the aim of the experiment. Par-

ticipants were asked to start articulating a continuation before 1200ms after the presentation of the context noun in the short condition and between 1600 to 2800ms following the context noun in the long condition. Therefore the same item appeared in four different conditions (nominative-short, nominative-long, accusative-short, and accusative-long). The 160 stimuli in the four conditions were distributed across four lists and assigned to participants using the Latin-square method.

Procedure

The experiment was conducted via the online experiment platform PCIbex (Zehr & Schwarz, 2018). The experiment consisted of two blocks. The long-condition trials were presented in the first block and the short-condition trials were presented in the second block. I separated the long-condition trials and the short-condition trials in different blocks, rather than randomly interleaving them because it was unlikely that participants could adjust to production deadlines randomly changing from one trial to the next. The long-condition block always preceded the short-condition block because piloting showed that the short condition was very challenging for most participants and it was easier for them to be familiarized with the speeded cloze task through the long condition, in order to be able to meet the strict production deadline in the short condition.

The fixed order of the blocks might raise the concern that practicing through the long block might make participants better at using argument roles in the short condition. Then, if the speeded cloze task shows role-sensitivity in the short condition although the EEG study did not, this could have been caused by the practicing in the long condition. I will later demonstrate that this is unlikely because there is no effect of practicing within each block.

In each trial, a fixation cross was presented in the center of the screen for 800ms, and then it was replaced with a context noun with a case marker. Participants were asked to produce a natural verb as a continuation to the presented context noun within the response time window. The response time windows were indicated in different ways in the two conditions, so that we could obtain responses in specific intervals. In the short condition, context nouns were presented for 1200ms and participants were asked to provide responses while the stimuli were being presented. In the long condition, context nouns were presented in a black font color for 1600ms, then they turned red and remained on the screen for the next 1200ms, and then disappeared. Participants were asked to produce responses while the font color was red, so that they do not produce responses before the response time window of the long condition. Their responses were recorded by PCIbex and were automatically sent to our server for data collection.

Data Processing

The recordings were manually transcribed and coded for onsets by the author using Praat (Boersma & Weenink, 2023), through both listening to the audio and visually inspecting the spectrograms. The coder had access to the context nouns but without case markers, in order to avoid bias in coding. After this process, each response was coded for role-appropriateness. Each pair of a context noun without case markers and a produced verb was classified into one of the three categories by the same author: (1) clearly more plausible with the nominative case (e.g., *bee* and *sting*) (2) clearly more plausible with the accusative case (e.g., *fly* and *swat*), or (3) no clear contrast between different cases (e.g., *cop* and *see*). The coder did not have access to the case marker on the context noun in order to avoid biasing. Subsequently,

each response was automatically coded according to the three categories. Responses produced with the preferred case were coded as role-specific and appropriate, which I henceforth refer to as role-specific. Other responses produced with the dispreferred case were coded as role-inappropriate, and the rest were coded as neutral.

Exclusion

First, I identified participants who failed to follow the instructions by inspecting their performance in the same 40 experimental items. I excluded participants who produced short condition responses in the long condition time window or vice versa in more than 50% of trials. Because participants can, in principle, complete the experiment by producing the same verb in different trials regardless of the context, they were explicitly instructed to avoid producing the same verb in many trials. In order to detect participants who did not follow this instruction, I counted the number of each verb produced by each participant. I excluded participants who produced the same three verbs in more than 50% of the trials. 19 participants were excluded by these two criteria, which left 54 participants.

Next, we excluded trials where non-words, non-verbs, or unrelated words were produced, or no responses were recorded at all. We also excluded trials where participants produced responses in the wrong time window (e.g. producing short condition responses in the 1600-2800ms long condition time window). We included short-condition responses that were produced before the beginning of the long condition time window, because many participants struggled to meet the deadline of the short condition. However, the short condition responses were on average still produced in the expected time window, and the results did not differ if we focus on the responses produced before the 1200ms deadline. (See below for more details.)

These exclusion processes left 6612 responses, which corresponds to 76.5% of all responses from the 54 participants.

3.4.2 Results

Reaction tiems

Reaction times are summarized in Table 6. We confirmed that short condition responses including those produced after the deadline (1200ms) but before the long condition time window (1600ms) were on average comparable in timing to the short condition of our EEG study. Since it is estimated that it takes about 350ms to start articulating verbs after completing lexical processing, on average the responses reflect activation up to 830ms after the presentation of the context noun. This is comparable to the short condition of our EEG experiment where participants had an 800ms interval between the onset of the context noun presentation and the target verb presentation.

Condition	Average RTs	Standard deviations
Long	2358.07	199.61
Short	1174.81	210.46
Short $<$ 1200ms	1019.67	124.30

Table 6: Average cloze reaction times (speech onset times) in milliseconds in the short condition and the long condition in ms.

Response categories

The proportions of responses in each category are summarized in Table 7. This excludes syntactically irreversible responses, such as intransitive verbs (28.9% of all the responses). We also excluded passives and causatives, because these types of responses require atypical case markings (2.3% of all the responses). As indicated in the table, more than two thirds of the responses had a clear bias for the roles of the context nouns and they were produced

following the more appropriate case marker (e.g., *Bee*-NOM *sting*-PRES). Combined with the neutral responses, plausible responses accounted for 94% of the responses, and role-inappropriate responses accounted for only 6% of the responses. Importantly, this pattern was consistent across the short and the long condition, as well as for the short condition responses that met the 1200ms deadline. We conducted Pearson's chi-squared tests for both the long and the short conditions, excluding the neutral responses. The tests showed that the frequency of role-specific and inappropriate responses was different in both conditions ($p < .001$).

Condition	Specific	Neutral	Inappropriate
Long	74.1%	22.2%	3.6%
Short	67.0%	27.0%	6.0%
Short < 1200ms	70.8%	24.0%	5.2%

Table 7: Proportion of role-specific, neutral, and role-inappropriate responses in the cloze task. $N = 2359, 2195, 1240$ for each row.

Table 8 shows the proportions of the role-inappropriate responses divided into the first half and the second half of each block. The differences between the first and second halves were much smaller than the differences between categories in Table 7. This suggests that people did not become significantly better at using argument roles in the short block because it was preceded by the long block.

Condition	First half	Second half
Long	3.8%	3.5%
Short	6.3%	5.8%

Table 8: Proportion of role-inappropriate responses in the cloze task, for the first and second halves within each block.

I also computed cloze probabilities for the specific target verbs used in the EEG study with different case markers. This analysis included syntactically irreversible responses, in

order to compute the cloze probabilities. As shown in Table 9, the target verbs were mostly produced in the appropriate context, and they were very rarely produced as reversals. A binomial generalized mixed-effects model revealed a significant effect of appropriateness ($\beta = 4.53, p < .001$). The same statistical model also revealed a significant interaction between SOA and appropriateness ($\beta = 1.89, p < .001$), showing that the difference between the target verb production in the appropriate and the inappropriate contexts was greater in the long condition.

Condition	Appropriate	Inappropriate
Long	19.0%	0.6%
Short	13.2%	1.7%
Short < 1200ms	15.6%	1.8%

Table 9: Cloze probability of the target verbs used in the EEG study. The left column shows the average cloze probability in the contexts where the target verbs are role-appropriate continuations, while the right column shows the average cloze probability in the contexts where the target verbs are role-inappropriate continuations. $N = 3430, 3182, 1772$ for each row.

3.4.3 Discussion for the cloze experiment

The cloze study showed that role-inappropriate responses are very rare in both the short condition and the long condition. The pattern did not differ between the analyses including or excluding the short-condition responses produced in the 1200-1600ms time window after the context presentation, which fell in between the two target response intervals. This shows that argument roles influence the cloze responses, even in a time window where N400 effects were not observed for the same stimuli.

It is important to note that 22.1% of the short condition trials had no responses or responses produced later than 1600ms, and they were excluded from the analyses above. While this is a relatively large proportion, it is not sufficient to explain the large gap between

role-appropriate responses and inappropriate responses.

3.5 Interim Summary

The analysis of the EEG data collected by Momma (2016) replicated the well-known absence of N400 effects in the short condition. Even though *sting* should be expected after *bee*-NOM but not after *bee*-ACC, there was no evidence for an N400 contrast at the verb. More importantly, N400 contrasts were found for the same pairs of stimuli in the long condition. This not only replicates Chow et al. (2018)'s findings but also extends them. This is because the stimuli were maximally simple in the experiment and the only difference between the short and the long condition is the additional 400ms added to the interval between the context noun and the target verb. This is stronger evidence that the absence of N400 effects in argument role reversals is specific to an earlier stage of processing, prior to the presentation of the target verb, and hence that N400 effects change dynamically as a function of time after the presentation of the context, but before the presentation of the target verb.

Previous EEG studies of argument role reversals have reported a late positivity for the reversed sentences compared to the canonical sentences (P600 effects). This was also the case for Chow et al. (2018), where they observed P600 effects both in the short and the long conditions. This has been considered to reflect the effort of semantic integration of the lexical item (e.g., Brouwer et al., 2017) or syntactic reanalysis (e.g., van Herten et al., 2005). Interestingly, significant P600 effects were not found in the current EEG data analysis either in the short condition or the long condition. Although I do not have a clear account for the absence of P600 effects in the EEG study, I suspect that this was due to the simplicity of Momma's stimuli, which might have resulted in relatively small integrative cost or revision

effort. There was little reason for participants in the current study to have any uncertainty over whether they had correctly understood the context.

While it was confirmed that N400 amplitude can be insensitive to argument roles at least in the short condition time window, in the speeded cloze task people produced role-appropriate verbs in the vast majority of cases both in the short condition and the long condition. There was also a large contrast in the cloze probability when we specifically analyzed the target verbs used in the EEG experiment. Importantly, the timing of the short condition was carefully matched between the cloze experiment and the EEG experiment. People had 800ms to process the context and activate likely continuations in the EEG experiment, while they took 830ms on average in the cloze experiment, considering the time required for post-lexical processes such as motor planning (350ms; Indefrey & Levelt, 2004). It was well known that people can give appropriate, role-sensitive completions in standard untimed cloze tasks. These have been used as a source of norming data for many prior EEG studies. What is novel in our findings is that the same role sensitivity is also found when speakers have much less time to respond, even when it is the same amount of time as in the EEG study. Thus the difference between N400 and cloze measures cannot be simply reduced to the contrast between online and offline comparisons, as has often been assumed.

3.5.1 Accounting for the mismatch between measures

The mismatch between the EEG and cloze studies is surprising given the Simple Activation Model, where the measures both reflect shared underlying pre-activation. This model was motivated by the parallelism between the measures observed previously, and the linking hypotheses for the different measures. It has been shown that the measures correlate with

each other in most cases (Kutas & Hillyard, 1984; DeLong, Urbach, & Kutas, 2005; Nieuwland et al., 2020). As discussed in the previous chapters, both N400 and cloze responses have been widely considered to reflect pre-activation.

However the EEG and speeded cloze data in the current study tell different stories about the influence of argument roles on lexical prediction. The absence of an N400 effect in argument role reversals, as replicated in the short conditions of our Japanese EEG study, has been taken to show that argument roles do not influence the pre-activation of upcoming lexical items, at least not initially. On the other hand, participants frequently produced role-appropriate responses and rarely produced role-inappropriate responses in the speeded cloze task, even if the responses were elicited in a comparable time window as the short condition of the EEG study. This suggests that argument roles impact the pre-activation of lexical representations, resulting in greater pre-activation for role-appropriate verbs than role-inappropriate verbs.

The puzzling mismatch between measures of pre-activation raises two questions: (i) Why do we find contrasts between different measures of prediction, even though the timing and the stimuli were carefully matched? (ii) Why do earlier and later presentations of continuations result in the presence/absence of N400 effects? In this section, I first address the possibility that people simply made different predictions in the cloze experiment and the EEG experiment, and reject the possibility. Subsequently, I propose two hypotheses where people pre-activate lexical candidates similarly across the tasks. However, one of them, the symmetric pre-activation hypothesis, posits that pre-activation is matched for role-appropriate and inappropriate candidates, whereas the asymmetric pre-activation hypothesis maintains that role-appropriate candidates are pre-activated more strongly compared to role-inappropriate

candidates.

Different predictions across tasks

It is in principle possible that the differences in the tasks made participants generate role-sensitive predictions in the speeded cloze task but role-insensitive predictions in the EEG study. One potentially important difference between the tasks is that participants are not presented with argument role reversal sentences in the (speeded) cloze studies, while most other paradigms involve the presentation of such sentences (EEG studies, eye-tracking studies, and the lexical decision and read-aloud studies in Chapter 5) [3]. Such explicit presentations of role-inappropriate candidates might have biased them to not rely on argument roles in pre-activating lexical candidates, because predictions based on argument roles are frequently disconfirmed by argument role reversals in these experiments. On the other hand, the (speeded) cloze task is free of such disconfirmation of predictions because participants are not presented with argument role reversals or any other types of anomalous sentences. In fact, some prior studies have shown that participants can make different predictions depending on their models of the experimental environment (e.g., how often they see semantically related lexical items: Lau, Holcomb, and Kuperberg (2013)).

We tested if the presentation of argument role reversals is the crucial factor accounting for the discrepancy between different measures in (Lee, Nakamura, Muller, & Phillips, 2022), where we interleaved speeded cloze trials with comprehension trials including anomalous sentences. Importantly, the participants did not know which type of trials they are in until

[3]Visual world paradigm is complicated regarding the presentations of argument role reversal sentences in that participants were not directly presented with those sentences, but they are presented with pictures of role-inappropriate objects while they listen to the contexts.

they are asked for a cloze response or a comprehension question in each trial. Even though argument role reversal sentences were presented, people produced as few role-inappropriate responses as in the speeded cloze tasks without explicit presentations of argument role reversals. Therefore, the presentation of argument role reversal sentences is unlikely to explain the greater role-sensitivity in the speeded cloze task compared with EEG studies.

Although the presentation of role-inappropriate candidates is not the only potential difference between the tasks, at this point, there does not seem to be a compelling account that can explain why people would generate role-sensitive predictions in the cloze tasks but not in the EEG studies.

Symmetric vs Asymmetric pre-activation

I propose two alternative hypotheses to account for the timing and task-based discrepancies in lexical prediction regarding argument roles: the symmetric pre-activation hypothesis and the asymmetric pre-activation hypothesis. The accounts share the notion that there are two stages of processing: (i) pre-activation of lexical representations, and (ii) the post-lexical combinatorial processes. The two hypotheses differ in which process is sensitive to argument roles. The symmetric pre-activation hypothesis attributes the role-insensitivity to the pre-activation of the verbs. Initially, role-inappropriate verbs are activated to a similar extent as role-appropriate verbs given the contexts (hence symmetric). N400 reflects this process relatively transparently, and shows no contrast between the role-appropriate and inappropriate verbs at least initially. However, there is a mechanism that monitors the plausibility of the utterance meaning representations where a pre-activated lexical candidate is incorporated into. This monitoring process rules out the role-inappropriate candidates in the cloze tasks

or in EEG experiments with sufficient time. This is responsible for the N400 contrasts in the long condition of the EEG study, and the role-sensitive productions in the speeded cloze study regardless of the timing.

On the other hand, pre-activation is sensitive to argument roles in the asymmetric pre-activation hypothesis and role-appropriate candidates are pre-activated more than role-inappropriate candidates (hence asymmetric). The utterance meaning representations are responsible for role-insensitivity, as some EEG studies on argument role reversals have argued (e.g., Kolk et al., 2003; Kim & Osterhout, 2005; Rabovsky et al., 2018). Under this account, pre-activation is sensitive to argument roles, and there is a large contrast between the pre-activation of role-appropriate and inappropriate verbs. The cloze task reflects this and hence responses are predominantly role-appropriate. However, once a role-inappropriate verb is presented and is retrieved (via bottom-up cues) in a comprehension task, it is as easy to incorporate it into the combinatorial semantic representations as role-appropriate verbs, and N400 amplitudes reflect such role-insensitive integration processes rather than lexical access (van Herten et al., 2005; Rabovsky et al., 2018). Moreover, the role-inappropriate verb might even trigger a heuristic interpretation ignoring the argument roles. For example, people might interpret *Bee*-ACC *sting.* as "A bee stings something." instead of "Something stings a bee." because the heuristic interpretation is way more plausible and attractive than the literal interpretation. The absence of N400 effects might reflect the lack of semantic anomaly in those cases (Kim & Osterhout, 2005; Li & Ettinger, 2023). Thus pre-activation is sensitive to argument roles in the asymmetric pre-activation hypothesis, but N400 amplitudes reflect role-insensitive combinatorial processes rather than pre-activation. Under this account, these combinatorial processes would need to become role-sensitive given additional

time.

Although the current EEG and speeded cloze experiments themselves do not provide strong evidence for or against either hypothesis, the contrast between the short and the long conditions in our EEG study is more straightforwardly consistent with the symmetric pre-activation hypothesis than the asymmetric pre-activation hypothesis. The Japanese EEG experiment controlled the time before the presentation of the target verb. Therefore, the N400 effects found in the long condition must be attributed to the processes that unfold before the presentation of the target verb rather than after it. While the symmetric pre-activation hypothesis attributes the N400 effects to the pre-activation accumulated before the presentation of the target verb, the asymmetric pre-activation hypothesis attributes them to the combinatorial processes following the presentation of the target verbs. Therefore, the timing contrast can be more straightforwardly explained by the symmetric pre-activation hypothesis.

Furthermore, the symmetric pre-activation hypothesis is more consistent with studies using other measures of pre-activation than the asymmetric pre-activation hypothesis. Consistent with argument role reversal literature in EEG, a few studies have reported that eye-tracking measures of pre-activation do not show sensitivity to argument roles, at least temporarily. Kukona et al. (2011) used a visual world paradigm (VWP) and presented a set of pictures including a picture of a role-appropriate candidate (e.g. thief) and a picture of a role-inappropriate candidate (e.g. cop) to participants as they listened to sentences such as (17). The participants looked at the picture of a cop as frequently as that of a thief before they heard *thief* in the sentence. Given the linking hypothesis for anticipatory looks in VWP introduced in Chapter 1, this suggests that the role-appropriate *thief* and the role-inappropriate

cop were equally pre-activated. Burnsky (2022) also found role-insensitivity in argument role reversals in eye movements while reading. Burnsky presented participants with sentences such as (18), and he did not find any difference between (18a) and (18b) in skipping rates and the first fixation times of *saved*. Since skipping rates and first fixation times are considered to reflect pre-activation, and since Burnsky found a difference between *saved* in go-past time, which is thought to reflect processes later than lexical access, this finding suggests that *saved* is pre-activated similarly in (18a) and (18b). These findings show patterns similar to N400 amplitudes and are consistent with the symmetric pre-activation hypothesis, if they indeed reveal lexical pre-activation. On the other hand, they are not straightforwardly consistent with the asymmetric pre-activation hypothesis, where pre-activation is sensitive to argument roles. The asymmetric pre-activation hypothesis has to explain the role-insensitivity of the eye-tracking measures despite the role-sensitive pre-activation, for example by revising the linking hypotheses.

(17) Toby arrested the thief.

(18) a. The parent saw which child the lifeguard had saved in the pool on Saturday.

b. The parent saw which lifeguard the child had saved in the pool on Saturday.

However, these findings from other measures may not be critical evidence against the asymmetric pre-activation hypothesis. First, while Kukona et al.'s study shows similar patterns to EEG studies on argument role reversals, the stimuli are different from those EEG studies or the current EEG and speeded cloze studies. Crucially, the target items in Kukona et al. are arguments, and the verbs are provided in the context, unlike the prior EEG studies. Although the exact contribution of verbs in the contexts is yet to be known, participants

could have generated different predictions in Kukona et al.'s study compared to the EEG studies. Second, there are some unexpected patterns found in Burnsky's eye-tracking study. Most EEG studies including the current study involve plausibility judgment immediately after the presentation of the stimuli. Typically, participants judge argument role reversal sentences to be implausible in about 80-95% of trials in these tasks (Kuperberg et al., 2003; Kolk et al., 2003; Hoeks et al., 2004; Kim & Osterhout, 2005; Chow et al., 2018; Liao et al., 2022). However, in Burnsky's study, about half of the argument role reversal sentences such as (18b) were judged to be plausible. This might suggest that participants did not comprehend the stimuli as accurately as in the EEG studies. Given these potential issues and that the literature on argument role reversals using these measures is much smaller than those in EEG, it is not clear how much they are consistent with the EEG findings on argument role reversals, and so these findings should be taken with caution.

In order to further sharpen the symmetric pre-activation hypothesis, I propose a computational model for this hypothesis and demonstrate that it successfully predicts the data pattern obtained in the two experiments. In this model, role-inappropriate candidates are activated just as much as role-appropriate candidates by the context. Once a candidate reaches a threshold, a monitoring mechanism that checks for appropriateness is triggered, and the role-inappropriate candidates are then detected and are inhibited. This account reduces the sensitivity to argument roles to the issue of how often role-incompatible candidates reach the threshold within a specific time limit to trigger the monitoring mechanism and to be successfully detected. Either additional time for activation or a lowered threshold to quickly produce a continuation would allow role-inappropriate candidates to reach the threshold.

The monitoring mechanism is related to the monitoring process in language production.

Levelt (1983) claimed that people monitor their own speech in speech production and repair the utterance when an anomaly is detected. This repair can be done not only on overt speech but also on internal speech that is planned but not yet articulated. The monitoring mechanism in lexical prediction can be considered as one form of such covert monitoring of speech in language production, where the plausibility of the planned utterance is checked.

3.6 Modeling

3.6.1 Model structures

I developed a computational model of pre-activation of lexical candidates, building on the Race model (Staub et al., 2015). Given a context, lexical items are activated in parallel, and the candidate lexical item whose pre-activation reaches a threshold first (i.e. the "winner" of the race) is produced as the cloze response. Because there is some amount of random variation in the activation, the most likely continuations are produced often but not all the time.

I take the basic race model and generalize and extend it to account for our findings in argument role reversals. I generalize the model to also account for the EEG studies with comprehension tasks. In comprehension tasks, lexical items are activated by context in the same way as in the cloze task, until a continuation is presented. N400 amplitudes are operationalized as the distance between the threshold and the amount of activation when the continuation is presented (Figure 29). More importantly, the model has a monitoring mechanism that inhibits the activation of inappropriate candidates that reached the threshold. Therefore the threshold in this model is considered as a trigger for the monitoring process.

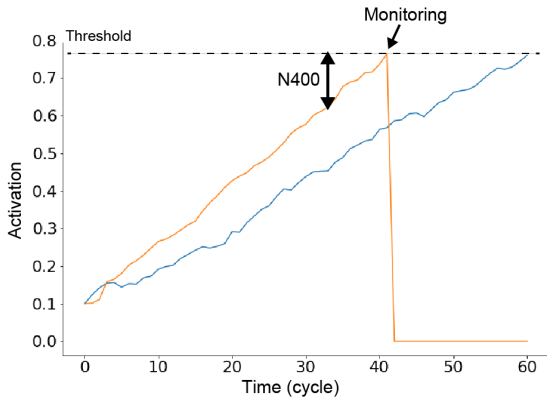


Figure 29: An illustration of a simulation of the model. I simulated step-by-step accumulation of activation, where the amount of activation in each step is determined by sampling from normal distributions. Once a candidate reaches the threshold, a monitoring process is triggered and the candidate is inhibited if it is inappropriate. N400 amplitudes are operationalized as the distance between the threshold and the amount of activation at the point of presentation of the continuation.

The structure of this model is almost identical to the step-by-step simulations in Section 2.5. In the model, each candidate's activation is updated in each 25ms-cycle. The amount of activation each candidate accumulates is randomly determined by sampling from normal distributions. While the σ parameter (i.e. variability) of the normal distributions is kept constant, each candidate has different values for the μ parameters (i.e. mean). More likely candidates have greater values for the μ parameters, so that they are more likely to accumulate activation in each step.

Once a candidate reaches the threshold, a monitoring process checks for its appropriateness in the context. When the winner is role-inappropriate, its activation is reduced to 0. In comprehension tasks, this process of activation and monitoring continues until a continuation is externally presented. On the other hand, in a cloze task the process ceases once a role-appropriate candidate reaches the threshold and passes the monitoring process. That candidate is produced as the cloze response.

In this model, the rare role-inappropriate responses in the speeded cloze task are considered as last resort cases, when no role-appropriate candidate reaches the threshold within the available response time window. In such cases, if there is a role-inappropriate candidate that has reached the threshold but is rejected through the monitoring process, the model may produce the role-inappropriate candidate rather than not producing anything. The detailed processes are described below and in Figure 30.[4]

While this model assumes identical activation processes for the speeded cloze task and for EEG studies with comprehension tasks, the difference between the tasks is implemented via the threshold for lexical activation. We propose that people set a lower threshold for the speeded cloze task compared to the comprehension task in the EEG studies due to the task demands of the speeded cloze task. Since participants are asked to provide a response within a specific time window in the speeded cloze task, they would try to have a lexical item that reaches the threshold before the end of the response time window. If they lower the threshold, there is more chance that they have some lexical item to produce within the response time window.

In our simulation, we focused on the contrast between the short condition of Momma (2016)'s EEG study and (i) the long condition of the same study, and (ii) the short condition of the speeded cloze study. This is because the absence of N400 effects in the short condition is consistent with prior studies, and the other two findings diverge from the general pattern and represent timing effects and task effects. We simulated 10000 races for five pairs of role-appropriate and inappropriate candidates with the same mu parameters

[4]Alternatively, role-inappropriate responses might be produced when they reach the threshold but are not detected by the monitoring mechanism. However, I will demonstrate in Section 4.3 that this view makes a wrong prediction about role-inappropriate response latencies.

($\mu = 0.024, 0.0215, 0.019, 0.0165, 0.014$) and the same sigma parameter ($\sigma = 0.01$). We set different thresholds for the comprehension task (1) and the production task (0.8).

In the comprehension task, the model was presented with either the best role-appropriate or inappropriate candidate (i.e. the one with the best μ parameters of 0.024) at 800 ms (i.e. 32nd cycle) or 1200 ms (i.e. 48th cycle) after the beginning of the race. The N400 amplitude was modeled by subtracting the activation of the presented candidate from the threshold of 1 at either timing. In the production task, the model keeps track of the role-inappropriate candidates that were rejected by the monitoring mechanism. If no role-appropriate candidate reached the threshold by the deadline but there is a role-inappropriate candidate that was rejected, the model produces the rejected candidate 25% of the time and does not produce anything in the remainder of the cases (Figure 30). This ratio was chosen so that the model does not provide responses in about the same proportion of trials as humans (c.f. there were 22.1% no-response or late-response trials in the short condition of the human cloze data). The model's response (or non-response) was recorded for each trial, and we computed the proportion of role-appropriate and role-inappropriate responses. All other parameters were determined to make the simulation match the human data in the current studies.

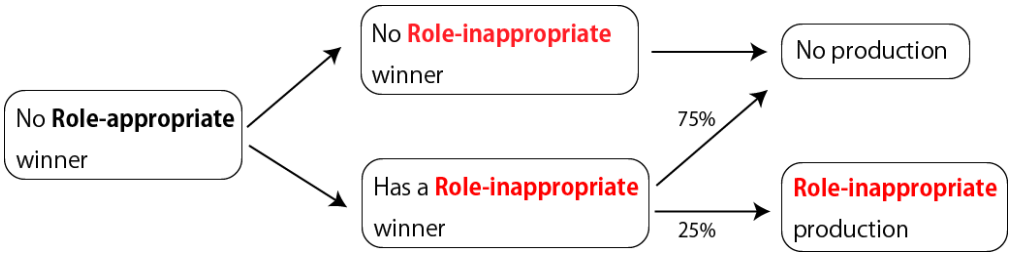
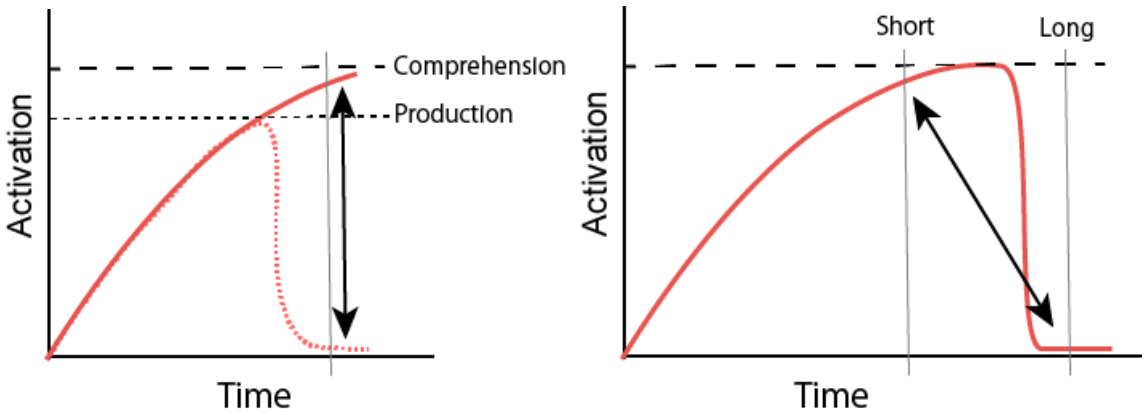


Figure 30: Illustration of the last resort production of role-inappropriate candidates.

The model expects that either additional time in the comprehension task or a low-

ered threshold in the production task should result in more successful inhibition of role-inappropriate candidates. As illustrated in Figure 31a, additional time for activation would allow role-incompatible candidates to reach the threshold and be detected and inhibited by the monitoring process. Similarly, Figure 31b shows that a lowered threshold in the production task also allows more successful inhibition of role-inappropriate candidates given the same amount of time.



(a) Timing **(b)** Threshold

Figure 31: Activation pattern of role-inappropriate candidates with (a) shorter vs longer time for activation, and (b) higher vs lower threshold for evaluation.

In this version of the model, competing items accumulate activation independent of one another, with no lateral inhibition between them, as also assumed by Staub et al. (2015). This contrasts with evidence for lateral inhibition mechanisms in Chapter 2, as well as in another speeded cloze study (Ness & Meltzer-Asscher, 2021b) and studies of lexical access in comprehension and production (Dahan et al., 2001; Chen & Mirman, 2012). The current model leaves aside this mechanism in the interest of simplicity of exposition. However, we also conducted simulations that incorporated a lateral inhibition mechanism, yielding comparable results (see Appendix 1).

3.6.2 Simulation results

We report the model-simulated N400 effects and the proportion of role-appropriate responses. Figure 32 shows the distance between the threshold and the activation of the presented candidates, which is assumed to be reflected in N400 amplitudes. The contrast between role-appropriate and inappropriate candidates was smaller for the short condition and greater for the long condition, successfully capturing the N400 patterns in our EEG experiment.

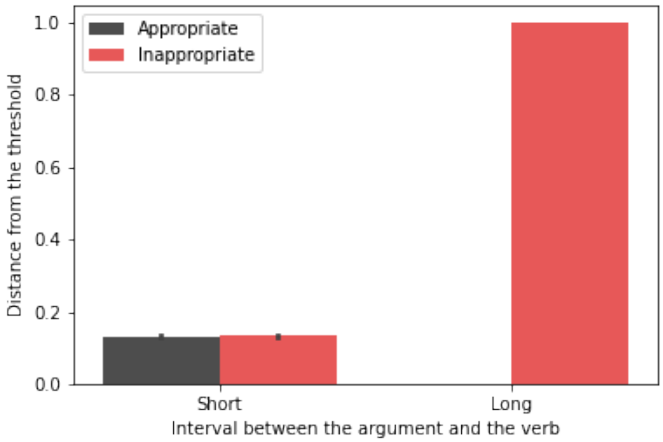


Figure 32: Simulated N400 amplitudes with shorter (left) vs. longer (right) time for activation (left).

In the production task, the responses generated by the model were overwhelmingly role-appropriate (93.2%) and role-inappropriate responses were rare (6.8%). No responses were produced in 17.5% of trials, which is comparable to the proportion in the human cloze data (22.1%). Thus the model also successfully predicted the human data pattern in our cloze task.

3.7 General Discussion

In the current studies, I investigated a potential case where different pre-activation measures do not align with each other. Many EEG studies have shown that N400 effects are absent in argument role reversals (Kolk et al., 2003; Kuperberg et al., 2003; Hoeks et al., 2004; Kim & Osterhout, 2005; Oishi & Tsutomu, 2010; Chow, Smith, et al., 2016; Chow et al., 2018; Liao et al., 2022), which suggests that argument roles do not influence lexical pre-activation. On the other hand, it is known that offline cloze probability is sensitive to argument roles (Chow, Smith, et al., 2016; Chow et al., 2018; Liao et al., 2022) and potentially in the speeded cloze task (Chow et al., 2015). Such a discrepancy between pre-activation measures could be caused by the difference between the measures or the tasks or by the difference in online and offline measures of prediction.

In order to identify the role of the task and timing manipulations, I reanalyzed an EEG study of Momma (2016) that used maximally simple Japanese stimuli and ran a speeded cloze experiment closely matched with it in terms of the time course. The experiments replicated the widespread finding that role reversals fail to elicit an N400 effect, when the verb appears shortly after the context argument(s). They also corroborated the finding that role reversals elicit an N400 effect when there is additional time between the context argument(s) and the verb. I also found clear evidence for role sensitivity, even without additional time, when pre-activation was measured via a speeded cloze task. Thus there are two closely related contrasts in argument role (in)sensitivity to be explained. One is about timing, and the other is about measures.

We offered two hypotheses to account for these contrasts that highlight the distinction

between the context-driven lexical activation and the integration of the lexical representations. One of them attributes the role-insensitivity to the lexical activation stage and the other attributes it to the integration stage. Our computational modeling offers a proof of concept demonstration for the former hypothesis where context-driven activation is initially insensitive to argument roles, but role insensitivity arises when role-inappropriate candidates reach a threshold, detected by an evaluation mechanism, and are successfully inhibited. This is more likely to occur when additional time is available for activation to accumulate towards a threshold in a comprehension task. It is also more likely when the threshold for evaluation is lowered due to the task demands in a speeded cloze paradigm.

3.7.1 Relationship to existing work: Experiments

The current EEG study strengthens the evidence from Chow et al. (2018) that additional time yields role sensitivity in EEG paradigms. An advantage of the simple Japanese sentences used in our study is that they made it possible to implement a "pure" timing manipulation, in which the short and long conditions differed only in the length of the time interval between the context noun and the target verb.

The findings are inconsistent with some existing accounts of the lack of N400 effects in role-reversal sentences. One approach attributes role insensitivity in the N400 component to a failure to parse the context fast enough to identify the argument roles of the items in the context. Our findings cast doubt upon this approach because we observed role-insensitivity in N400s following exceedingly simple contexts, which speakers are unlikely to struggle to parse. Furthermore, I found role sensitivity in responses to the same context, on the same time scale, when the task involves a speeded cloze response. This suggests that argument

roles must be identifiable by the time when N400 is elicited or cloze responses are planned.

Another family of approaches attributes the role insensitivity in the N400 component to consideration of unsupported but highly attractive verb-argument pairings. A variant of this approach, proposed by Kuperberg (2016), proposes that argument role information in the context is selectively ignored in prediction processes, specifically in cases where ignoring the role information more readily yields a continuation than taking the role information into account. For example, just the set or a sequence of {waitress, customer} may provide constraining predictions for the continuation, because it is very likely for a waitress to serve a customer. On the other hand, the full set of cues {customer-agent, waitress-patient} may result in less constraining prediction about the continuation because there may not be very likely events involving customer-as-agent and waitress-as-patient. In that situation people might rely more on the mere set or sequence of the arguments, excluding their roles, because they result in more constraining predictions and hence are more reliable. This is an unlikely account of the role insensitivity observed in the Japanese materials, where the context consists of a single noun and a case marker. It seems unlikely that people can generate more constraining predictions given the single arguments without roles such as *bee*, than the combination of arguments and roles such as *bee*-as-agent.

3.7.2 Relationship to existing work: Modeling

Accounts of the absence of N400 effects

We highlighted two kinds of approaches to the contrasting profiles of role (in)sensitivity in our studies. We used an implemented computational model to simulate one of these approaches, in which initial activations are role insensitive, and where role sensitivity is the result of a

serial evaluation process. Our simulation serves as a proof of concept demonstration of that account, rather than as evidence that it is the preferable account. However, this simulation is consistent with existing accounts that assume initial role-insensitive pre-activation of lexical candidates.

Brouwer, Fitz, and Hoeks (2012); Brouwer et al. (2017) attribute insensitivity to lexical activation caused by simple association and scenario-based world knowledge. That is, *sting* is activated by *bee*-ACC because (i) *bee* and *sting* are associated, and (ii) *sting* is likely to appear in a scenario where *bee* is involved. The role-sensitive evaluation mechanism in our model can be understood as an additional process that constrains this association-based activation.

A related but slightly different account by Chow, Momma, et al. (2016) attributes role-insensitivity of N400 to event memory search. They suggest that lexical prediction requires retrieval of likely events from the long-term memory, and that argument roles might not be able to be used for retrieval of event memories. As one possible mechanism for this, they suggest that thematic roles might be encoded in event-specific manners such as stinger or stingee rather than abstract thematic roles such as agent or patient. With such memory representations, people can only search their memory using the arguments without roles as cues. They argue that events retrieved by such a role-insensitive memory search are rejected in some indirect manners, which takes some additional time. Our model can be taken as an extension of their suggestion that provides a concrete mechanism for this rejection process and simultaneously accounts for the cloze data.

These accounts are similar in that they attribute the lack of N400 contrasts between role-appropriate and inappropriate items to role-insensitive mechanisms of pre-activation. Furthermore, Liao et al. (2022) argue that both simple association and event memory search

contribute to the role-insensitive pre-activation but with different timing profiles.

The monitoring model and pre-updating

Our assumption of a serial evaluation process that leads role-inappropriate candidates to be inhibited looks like a departure from other existing models, but we see a number of precedents. Although accounts of prediction in language processing tend to focus on the parallel activation of multiple promising candidates, this cannot be the whole story. There must be additional steps that apply to specific individual candidates following the parallel activation.

In production, Staub et al. (2015)'s race model assumes the notion of an activation threshold, corresponding to the point at which a lexical candidate has accumulated sufficient activation for it to be ready for phonological processing. In this approach, the first item to reach the threshold is the one that gets articulated. The only addition in our account here is the suggestion that reaching this threshold also triggers an evaluation (or monitoring) process, which is a standard component of production models (e.g. Levelt, 1989).

In comprehension, too, simply pre-activating multiple lexical continuations cannot be enough. There must eventually be a point at which an individual item is integrated into the representation of the context. This could happen at the point when the actual incoming word is presented. But there is also interesting evidence that this process sometimes occurs before the presentation of continuations, in situations in which the comprehender has a high degree of confidence in an individual predicted continuation. Such early integration has been called pre-updating (e.g., Lau et al., 2013; Ness & Meltzer-Asscher, 2018, 2021a). The prior studies have found P600 effects at words preceding highly predicted continuations,

suggesting that the continuation is integrated before its presentation. Our model is consistent with this literature, but instead of assuming that a lexical item is sent to the integrative process immediately after reaching the threshold, the model has an additional mechanism that evaluates the plausibility of the lexical item in the semantic representation of the sentence.

The parallelism of our model and pre-updating makes an empirical prediction that there should be late positivity in the context noun of the plausible-long condition of the EEG study compared to that of the plausible-short condition, because additional time would allow role-appropriate candidates to reach the threshold and to be integrated into the sentential semantic representations. However, our current dataset cannot assess this prediction because if additional time allows more pre-updating, most of them should occur 800-1200ms after the presentation of the context noun. We do not have EEG data for this time window for the short condition, because participants had already been presented with the target verb by the time. Therefore, examination of this empirical prediction is left for future studies.

There is in fact a recent finding that supports the existence of the common monitoring mechanism in the speeded cloze tasks and the comprehension tasks in EEG studies. Lee and Phillips (2023) observed N400 effects for argument role reversals in an EEG experiment with comprehension tasks when speeded cloze trials were interleaved. Two points should be highlighted about their experiment. First, participants did not know if they were in a cloze trial or a comprehension trial until they were asked to provide a continuation or the sentence ended. Second, they observed the N400 effects in the comprehension trials, where they used the same stimuli as a prior EEG study on argument role reversals. This is consistent with the monitoring hypothesis that attributes the differences between the cloze studies and the EEG studies to the differences in the threshold. The interleaved cloze trials would generally lower

the threshold for the monitoring process because they did not know which type of trials they are in. Then role-inappropriate candidates in the comprehension trials, as well as the cloze trials, would be detected and inhibited by the monitoring mechanism. Thus the monitoring hypothesis seems to be supported by this study.

In the current study, we investigated potential misalignment of pre-activation measures in argument role reversals. The closely-matched EEG and cloze studies replicated the N400 amplitudes' insensitivity to argument roles in the short condition of the EEG study, but also showed that people show sensitivity to argument roles (i) with additional time in the same EEG study or (ii) at the same moment in a cloze paradigm. This raises a question of why different measures of pre-activation or different timing result in different sensitivity to argument roles. I proposed two hypotheses to account for the mismatch in pre-activation measures that attribute role insensitivity to different processes of lexical and semantic processing. In the symmetric pre-activation hypothesis, N400 amplitudes reflect the matched pre-activation for role-appropriate and inappropriate candidates, while cloze responses are mostly role-appropriate because role-inappropriate candidates are inhibited by a monitoring mechanism. In the asymmetric pre-activation hypothesis, cloze responses reflect the greater pre-activation for role-appropriate candidates, but later integrative processes are insensitive argument roles, which is reflected in N400 amplitudes.

While the symmetric pre-activation hypothesis has some advantage over the asymmetric pre-activation hypothesis, there is no conclusive evidence for either hypothesis at this point. In the following chapters, I seek evidence to evaluate the two hypotheses. In Chapter 4, I will show that role-inappropriate candidates are activated to some extent in the speeded cloze

task. In Chapter 5, I will show through other pre-activation measures that role-inappropriate candidates can be as pre-activated as role-appropriate candidates.

4 Cloze RT analyses to evaluate the role-inappropriate activation

In the previous chapter, the Japanese EEG and cloze studies showed that different pre-activation measures do not align with each other in argument role reversals, even when the time courses and stimuli are matched. In the EEG study, N400 amplitudes for the same verbs as role-appropriate continuations and role-inappropriate continuations were matched when the verbs were presented with a short delay after the presentation of the context. Under widespread assumptions about the N400, this suggests that the role-appropriate and inappropriate verbs are at least temporarily pre-activated to the same extent, despite the apparent contrast in their predictability. On the other hand, in the speeded cloze task, role-appropriate responses were much more frequent than role-inappropriate responses, suggesting that they are differently activated. This was the case even when participants were urged to produce cloze responses in the same time window as in the EEG study where N400 amplitudes were matched for role-appropriate and inappropriate continuations. These findings in the two pre-activation measures are consistent with the large prior literature in each measure (EEG: Kolk et al., 2003; Kuperberg et al., 2003; Hoeks et al., 2004; Kim & Osterhout, 2005; Oishi & Tsutomu, 2010; Chow, Smith, et al., 2016; Chow et al., 2018; Liao et al., 2022; online and offline cloze task: Chow et al., 2015; Chow, Smith, et al., 2016; Chow et al., 2018; Liao et al., 2022). They also provide strong evidence that the discrepancy between cloze probabilities and N400 amplitudes in argument role reversals cannot be attributed to the timing differences, as has been suggested previously (e.g., Chow et al., 2018).

I proposed two competing hypotheses regarding the apparently inconsistent findings from

the EEG studies and the cloze studies. One of the two hypotheses is the asymmetric pre-activation hypothesis. Under this hypothesis, the pre-activation of lexical candidates is not sensitive to argument roles. While N400 amplitudes relatively transparently reflect the matched pre-activation for role-appropriate and inappropriate candidates, role-inappropriate responses in the cloze task are prevented by a mechanism that monitors the plausibility of the utterance. The other hypothesis is the asymmetric pre-activation hypothesis, where the pre-activation process is sensitive to argument roles. Under this hypothesis, cloze responses reflect such role-sensitive pre-activation, but N400 amplitudes reflect later integrative processes that are not sensitive to argument roles.

As a first step in evaluating the two hypotheses, in this chapter, I demonstrate that the pre-activation mechanisms are not completely role-sensitive and that there is at least some pre-activation for role-inappropriate candidates. While this is fully consistent with the symmetric pre-activation hypothesis, it is inconsistent with a strong version of the asymmetric pre-activation hypothesis, according to which there is little or no activation for role-inappropriate candidates.

As a different approach in this chapter, I exploit the between-context variability in the availability of role-inappropriate candidates to investigate the role-inappropriate pre-activation. If role-inappropriate candidates are pre-activated, some contexts may have more tempting and stronger role-inappropriate candidates with high pre-activation rates ("lures"; e.g.,*tip* for (19)), while others may not have salient role-inappropriate candidates (e.g.,(20)). I conducted two speeded cloze experiments with English argument role reversal stimuli, taken from prior EEG studies (Chow, Smith, et al., 2016; Ehrenhofer et al., 2019) and demonstrate that such between-context variability in the availability of role-inappropriate candidates in-

fluences (i) the role-appropriate RTs and (ii) the proportion of role-inappropriate productions, suggesting that they are in fact pre-activated.

(19) The restaurant owner forgot which customer the waitress had ...

(20) The parent saw which child the lifeguard had ...

First, pre-activation of role-inappropriate candidates should influence the production latencies of role-appropriate responses. As demonstrated in Chapter 2, pre-activated lexical candidates appear to inhibit each other. Therefore, role-inappropriate candidates should impose inhibition on the role-appropriate candidates and should slow down the responses, even if the role-inappropriate candidates are not produced. I demonstrate that contexts that are estimated to have more pre-activation for role-inappropriate candidates have slower role-appropriate RTs compared to contexts without tempting role-inappropriate candidates. This provides evidence that role-inappropriate candidates are activated to the extent that they can inhibit role-appropriate candidates and slow down the role-appropriate responses. Second, if role-inappropriate candidates are pre-activated such that they can win races at least occasionally, contexts with stronger role-inappropriate candidates should have more role-inappropriate productions. I demonstrate that this pattern is observed in the speeded cloze experiments. Combining these two types of evidence from context-wise variability in the availability of role-inappropriate candidates, I argue that role-inappropriate candidates are pre-activated to the extent that they can inhibit role-appropriate candidates and can win races at least occasionally.

4.1 Experiment 4-1: Speeded cloze task with Chow et al. stimuli

The aim of this study was to test if there is some degree of role-insensitivity in pre-activation and if role-inappropriate candidates are pre-activated. However, the fact that role-inappropriate responses are rare makes it difficult to test this with the available speeded cloze data. Role-inappropriate responses accounted for about 6% of the short condition responses in the previous Japanese speeded cloze experiment (Chapter 3). These 6% responses might reflect a small but general pre-activation for role-inappropriate candidates. For example, a role-inappropriate candidate with 3% cloze probability might win 3% of races just like a role-appropriate candidate with the same cloze probability. Alternatively, this could mean that they are highly pre-activated, but in 6% of trials they are produced despite the monitoring mechanism. These accounts of role-inappropriate productions agree that role-inappropriate candidates participate in basically every race, even when role-appropriate candidates are produced. They are not produced frequently, either because their pre-activations do not reach the threshold before their competitors, or because the productions are prevented by a monitoring mechanism. However, it is also possible that role-inappropriate candidates are generally not considered but that there are 6% of exceptional trials where the role-inappropriate candidates are pre-activated and produced. For example, participants might misinterpret the context (21a) as (21b) and produce *served* in some trials. Given the small number of trials where role-inappropriate responses are produced, it is not clear whether these role-inappropriate responses in fact reflect sufficient pre-activation for production, or whether they reflect zero or minimal pre-activation but with occasional errors outside of the normal race process.

In order to test if there is general pre-activation for role-inappropriate candidates, I use

between-context variability of the strength of role-inappropriate lures and demonstrate that lures systematically affect cloze responses. If role-inappropriate candidates are pre-activated, some contexts may have tempting but role-inappropriate candidates (e.g., *tip* for (21a), while others may not have such tempting role-inappropriate candidates (e.g., (22a)). If there is pre-activation for role-inappropriate candidates, those candidates should influence cloze responses in multiple ways.

(21) a. The restaurant owner forgot which customer the waitress had ...

b. The restaurant owner forgot which waitress the customer had ...

One way in which the role-inappropriate candidates can influence cloze responses is through lateral inhibition. Simulations in Chapter 2 provided evidence for lateral inhibition between pre-activated candidates. The original race simulations by Staub et al. (2015) failed to capture the slow low-cloze responses observed in their speeded cloze data. I demonstrated that such patterns can be explained by an additional assumption of lateral inhibition in the pre-activation mechanisms, which has independent support in the bottom-up lexical access literature. Given this finding, if role-inappropriate candidates are pre-activated, role-inappropriate lures should inhibit other competitors. In particular, role-appropriate responses should be produced more slowly if they are competing with stronger lures compared to role-appropriate candidates competing with weaker lures. Therefore, we can test for pre-activation of role-incompatible candidates via inhibitory effects on role-compatible responses, even if they are rarely uttered. Such inhibitory effects on role-appropriate RTs provide strong evidence that role-appropriate candidates are generally participating in races.

Another way to test for pre-activation of role-inappropriate candidates is through the

proportion of role-inappropriate responses that are produced. If the role-inappropriate responses in the cloze task reflect occasions on which they win races, there should be more role-inappropriate responses in contexts with stronger lures.

I propose that the strength of lures in a context can be estimated from the strength of the same lexical items as role-appropriate candidates in the context with the alternative word order (henceforth the "alternative context"). When a lexical item is strong as a role-appropriate competitor, presumably it is also strong as a role-inappropriate competitor in a context with the alternative word order (henceforth the "alternative context"). Then, when a context has strong role-inappropriate lures, it is likely to have strong role-inappropriate competitors in the alternative context. For example, the context (21a) seems to have tempting lures such as *tip*, and they are competitive role-appropriate candidates for (21b). On the other hand, there are no tempting role-inappropriate candidates for (22a), presumably because there is no salient role-appropriate candidates for (22b).

Based on the findings from Chapter 2 that the average of all RTs given the context can be used as a context-level measure of the availability of candidates, the lure strengths of each context can be estimated by mean RTs for all role-appropriate responses in the alternative contexts. For example, the lure strengths in (21a) can be estimated based on the mean RTs of role-appropriate responses in (21b). The difference in the availability of tempting lures between (21a) and (22a) can be captured by the difference in the mean RTs for role-appropriate responses between (21b) and (22b).

(22) a. The parent saw which child the lifeguard had ...

b. The parent saw which lifeguard the child had ...

One might be concerned that if pre-activation is role-insensitive and if there is a monitoring mechanism, as proposed in Chapter 3, cloze measures might not be able to reveal the availability of candidates, because role-appropriate RTs would be skewed by the monitoring processes. However, even in such cases, mean cloze RTs are still useful context-level measures. Contexts with stronger role-appropriate candidates are still associated with on average faster role-appropriate responses, because stronger role-appropriate candidates can win quickly and also prevent role-inappropriate candidates from winning races. Therefore, even if pre-activation might be role-insensitive, and there might be a monitoring mechanism, cloze RTs can still be used to estimate context-wise variability in the strength of role-appropriate candidates or lures.

In Experiment 4-1, I tested the pre-activation of role-inappropriate candidates via the context-wise variability in the lure strengths. I collected speeded cloze data using Chow, Smith, et al. (2016)'s stimuli. Chow, Smith, et al. conducted an EEG study where they used sentence pairs such as (23) and found that there is no contrast in N400 amplitudes for *served*. I obtained the lure strength for each context as the mean RTs of the alternative context, and tested if that covaries with the RTs of role-appropriate responses and the proportion of role-inappropriate responses. The results show that role-inappropriate candidates are in fact pre-activated.

(23) a. The restaurant owner forgot which customer the waitress had served during dinner yesterday.

b. The restaurant owner forgot which waitress the customer had served during dinner yesterday.

4.1.1 Methods

Participants

84 native American English speakers were recruited online on the crowdsourcing platform Amazon Mechanical Turk. Participants received monetary compensation. Five participants' data were not analyzed due to missing or unclear recordings, or not performing the task. Additionally, eleven participants were excluded from the data due to unsatisfactory performance as described in the Exclusion section below. This left 68 participants for analysis (36 females and 32 males; average age = 41.0).

Materials

We used 60 pairs of experimental items from Chow, Smith, et al. (2016) such as (24). Unlike the EEG experiment, each stimulus ended at the word preceding the target verb. Participants instead provided a word that was a natural continuation of the presented context as quickly as possible. Therefore, the same experimental item appeared in two different word orders (e.g., (21)). The 60 stimuli in the two word orders were distributed across two lists and assigned to participants using the Latin Square method. Additionally, 70 fillers were created based on the filler items used in Chow, Smith, et al. (2016). The original sentences were cut off in the middle so that participants could produce continuations.

Procedure

The experiment was conducted via the online experiment platform PCIbex (Zehr & Schwarz, 2018). In each trial, a fixation cross was presented on the center of the screen for 800ms.

Each word of the sentence fragment appeared word-by-word in black font color on the center of the screen. Each word was presented for 300ms and was followed by a blank screen presented for 230ms, which were the same as Chow, Smith, et al. (2016). Therefore the stimulus onset asynchrony (SOA) was 530ms. Participants were instructed to articulate a natural word as a continuation as quickly as possible after the final word of the sentence fragment was presented, which was indicated by red font color. The audio was recorded for 3000ms following the presentation of the final word, and then a text indicating the end of the trial was presented. The responses were recorded by PCIbex and were automatically sent to a server for data collection.

Data Processing

The recordings were first automatically transcribed and coded for onsets using the TransCloze pipeline developed by the author. This pipeline uses the Google Cloud Speech-to-Text API for automatic transcription of the speech data and Chronset (Roux, Armstrong, & Carreiras, 2017) for automatic detection of speech onsets. The transcriptions and the detected onsets were then manually corrected by the author using Praat (Boersma & Weenink, 2023), through both listening to the audio and visually inspecting the spectrograms. The coder had access to the arguments of the produced verbs, but the order of the arguments was kept constant for both sentence fragments of each pair, regardless of the actual word order in the experiment. For example, the coder was presented with {waitress, customer} for both (21a) and (21b). This helped the coder transcribe each response with contextual information but without biases based on argument roles.

Subsequently, an undergraduate RA whose native language is American English coded

each response. In this process, ungrammatical responses were flagged and removed from the further coding processes. Then each of the remaining responses was presented with the exact context in which it was produced as a continuation and the alternative context. The coder checked each response and judged if it was clearly more plausible in either of the two contexts, whether it was equally plausible in the two contexts, or whether it was not plausible in either context. In order to avoid biases, the coder did not know which of the two contexts was the actual context in which the response was produced. Finally, each response was then automatically coded based on the manual coding in the following way. Responses produced with the (clearly) more plausible word order were coded as role-specifically appropriate. I will refer to this category as role-specific for simplicity. Other responses produced with the (clearly) less plausible word order were coded as role-inappropriate. Responses that were equally plausible in either word order were coded as neutral. Responses that were not plausible in either word order were flagged for the exclusion process described below and were later removed from analyses.

Exclusion

I identified participants who failed to follow the instructions by inspecting their performance. I counted trials with no response, non-word responses, ungrammatical responses, or responses that are implausible in either word order. The trials where participants articulated the context word(s) were also flagged as anomalous trials. This was because the cloze reaction times might be affected by the articulation of those context words and participants were instructed to avoid doing that.[5]

[5]Since recording started when the final words of the contexts were presented, we cannot detect the cases

All participants who had 33% or more of those anomalous trials were excluded. Because participants can, in principle, complete the experiment by producing the same verb in different trials, regardless of the context, they were explicitly instructed to avoid producing the same verb in many trials. In order to detect participants who did not follow this instruction, I counted the number of each continuation produced by each participant. I excluded participants who produced the same continuation in more than 25% of trials. Eleven participants were excluded following these criteria, leaving 68 participants. Finally, I excluded individual anomalous trials for the remaining participants. These exclusion processes left 3337 responses, which corresponds to 81.8% of all responses from the 68 participants.

4.1.2 Results

All the statistical analyses below used linear mixed-effects models with the lmer function or binomial generalized linear mixed-effects models with the glmer function of the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) and lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2017) of the statistical analysis software R (R Core Team, 2020). I included the maximal random effects structure for subjects and items that converged. The fixed effects were different for each analysis, but they were all centered on their means.

Overall proportions of the categories

59.2% of the responses were role-specific, 26.5% were neutral, and 14.3% were role-inappropriate.

where participants articulated the context words before that. However, those cases are less problematic because such early articulation would have a relatively small effect on the production latencies of the cloze responses. Moreover, there were very few cases of (detected) articulation from participants who were included in the analyses, while some removed participants did that in most trials. Therefore, I assume that the articulation of the contexts has minimal effect on the cloze RTs

Role-specific and neutral responses (role-appropriate responses) accounted for the majority of the responses (85.7%). While there were more role-inappropriate responses in this experiment than in the previous Japanese speeded cloze study, the overall pattern that most responses were role-appropriate was replicated.

Context-wise variability in RTs of role-appropriate responses

Figure 33 illustrates the relationship between the role-appropriate RTs and the mean RTs of all role-appropriate responses in the alternative contexts (henceforth the alternative mean RTs). I conducted a statistical analysis using a linear mixed-effects model of RTs of role-appropriate responses in individual trials. In addition to the alternative mean RTs, I included the cloze probability of each response and its interaction with the alternative mean RTs. This is to account for the within-context variability of RTs, which cannot be accounted for by the alternative mean RTs. The analysis revealed a significant effect of the alternative mean RT ($\beta = -0.62, p < .001$), suggesting that stronger lures (shorter alternative mean RTs) resulted in slower role-appropriate responses. The black line in Figure 33 shows the coefficient of this fixed effect. There was also a significant effect of the cloze probability ($\beta = -0.61, p < .001$), replicating the negative correlation between cloze RT and cloze probability (Staub et al., 2015). The interaction between the alternative mean RT and the cloze probability was not significant ($\beta = -0.28, p = 0.48$).

Context-wise variability in the role-inappropriate responses

The proportions of role-inappropriate responses are plotted in Figure 34 as the color of each dot. I also plotted the mean RTs of the role-appropriate responses produced in each

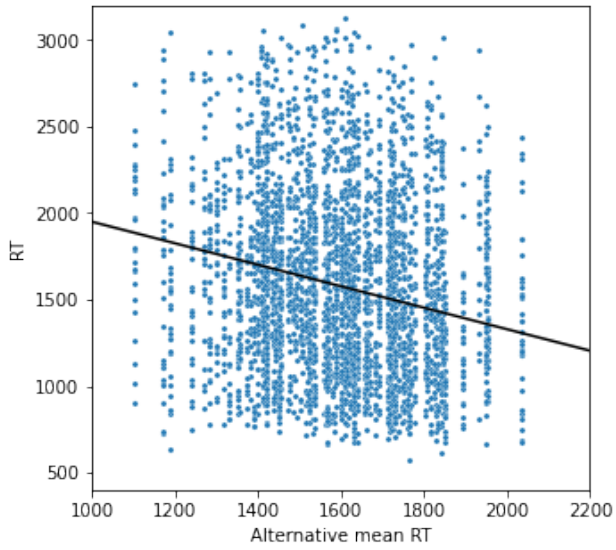


Figure 33: The relationship between alternative mean RTs (x-axis) and role-appropriate RTs (y-axis) for Experiment 4-1.

context, which is another type of context-wise variability that is expected to influence the production of role-inappropriate responses. If there are stronger role-appropriate candidates, there should be fewer role-inappropriate responses, because the role-inappropriate candidates should be less able to win races. The x-axis shows the mean RT of the role-appropriate responses produced in each context, and the y-axis shows the mean RT of role-appropriate responses produced in its alternative context. The color of the dots indicates the proportion of role-inappropriate responses. As can be seen in the figure, there were more role-inappropriate responses in contexts with shorter alternative mean RTs (i.e. stronger lures), indicated by green to orange dots. In general, there were more role-inappropriate responses in the contexts with longer mean RTs (i.e. stronger role-appropriate candidates).

A statistical analysis using a binomial generalized linear mixed-effects model of the role-inappropriateness of each response revealed significant effects of the mean RTs of the context itself ($\beta = 2.97, p < .001$) and the mean RTs of the alternative context ($\beta = -4.39, p < .001$).

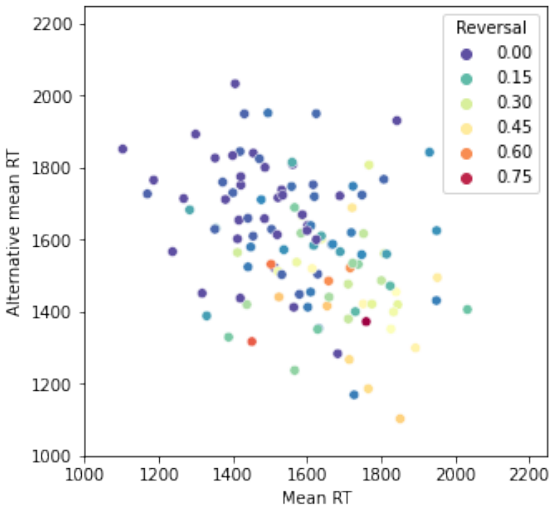


Figure 34: The relationship between mean RT (x-axis), alternative mean RT (y-axis), and the proportion of role-inappropriate responses (color).

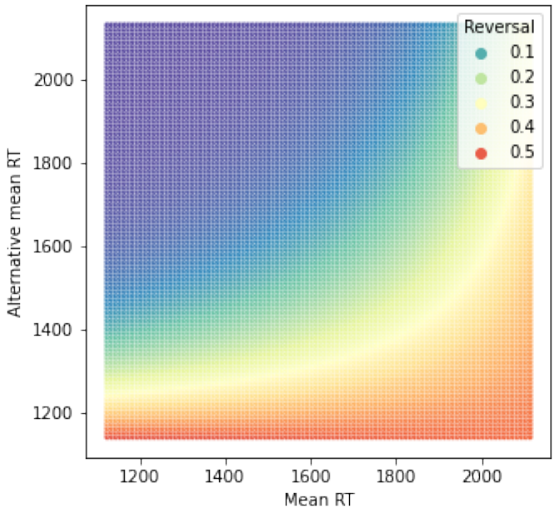


Figure 35: An illustration of the binomial generalized linear mixed-effects model's prediction. The axes and the color are identical to Figure 34.

These suggest that stronger lures or weaker role-appropriate competitors result in more production of role-inappropriate responses. Additionally, there was an interaction between the two effects ($\beta = 7.28, p < .001$). Figure 35 illustrates the model prediction to facilitate the interpretation of the interaction. There seems to be a consistent pattern showing that contexts with stronger role-inappropriate candidates are associated with more role-inappropriate responses, as expected. Contexts with weaker role-appropriate candidates are also associated with more role-inappropriate responses, which was also expected. However, this pattern regarding the role-appropriate candidates is reversed in the data with strong lures (i.e. the bottom of the figure). I suspect that this arises from the lack of data points with both short mean RTs and alternative mean RTs (the bottom-left corner of Figure 35).

4.1.3 Discussion

The results showed that contexts with stronger lures have slower role-appropriate responses and more role-inappropriate responses. These patterns can be straightforwardly explained if role-inappropriate candidates are pre-activated enough to inhibit role-appropriate candidates or to be produced occasionally. In particular, the analysis of role-appropriate RTs provides a new kind of evidence for role-insensitive pre-activation, because it shows that role-appropriate responses, which account for the majority of the cloze responses, are nevertheless affected by the pre-activation of their role-inappropriate competitors. This suggests that role-inappropriate candidates are activated even in trials where they are not produced.

However, the finding about the impact of role-appropriate RTs should be taken with caution, because the observed pattern could have been caused by the specific choice of stimuli. In most of Chow, Smith, et al. (2016)'s stimuli pairs, only one of the paired contexts had strong role-appropriate candidates. For example, *saved* is a role-appropriate response for (22a) with a high cloze probability and short RTs (cloze probability: 77.4%, mean RT for the response: 1086ms). On the other hand, in its alternative context (22b), the fastest response produced by more than three participants is *seen* (cloze probability: 16.7%, mean RT for the response: 1808ms). It is indeed difficult to create stimuli that have strong role-appropriate candidates with either word order. In fact, as can be seen from Figure 34, there are no contexts that have both short mean role-appropriate RTs and alternative mean RTs (i.e. no dots in the bottom left corner). Given this, role-appropriate RTs might have been slower when there are strong lures not because the lures inhibited the appropriate candidates, but because there were simply no strong role-appropriate candidates.

4.2 Experiment 4-2: Speeded cloze task with Ehrenhofer et al. stimuli

In order to address the possibility that one of the findings in Experiment 4-1 was caused by the nature of the stimuli rather than by the pre-activation of role-inappropriate candidates, I replicated these findings with a different set of stimuli. Specifically, I used context pairs with strong role-appropriate continuations in either word order created by Ehrenhofer et al. (2019). Ehrenhofer et al.'s experiment was very similar to Chow, Smith, et al. (2016)'s study, both in terms of the design and the stimuli, but they used context pairs that had matched offline modal cloze probabilities. For example, (24a) has a highly predictable continuation of *ridden* and (24b) also has *gored*. Although the modal cloze probabilities might not provide the best estimates of the availability of candidates (see Chapter 2), these stimuli can have both strong role-appropriate candidates and strong lures. A speeded cloze study with such stimuli can therefore provide better evidence for role-inappropriate pre-activation than Experiment 4-2.

(24) a. The cattle rancher remembered which bull the cowboy had ...

b. The cattle rancher remembered which cowboy the bull had ...

It should be noted that Ehrenhofer et al.' study was different from other EEG studies on argument role reversals including Chow, Smith, et al.'s study in that they found N400 effects for the role-inappropriate verbs, which was also confirmed by another experiment. I will discuss this in more detail in Chapter 6.

4.2.1 Methods

Participants

80 native American English speakers were recruited in the same way as in the previous experiment. Three participants' data were not successfully uploaded to the server and were removed from analyses. Five participants' data were not analyzed due to multiple submissions by the same person. Four additional participants were removed for not following the instructions. Finally, three participants were excluded from the data due to unsatisfactory performance as described in the Exclusion section below. This left 64 participants for analysis (32 females and 32 males; average age = 40.9).

Materials

We used 60 pairs of experimental items from Ehrenhofer et al. (2019) such as (24). One noun phrase in an original pair was replaced with a less sensitive expression with minimal change in the meanings of the sentence. Ehrenhofer et al. (2019) used four experiment items that were also used in Chow, Smith, et al. (2016), and I included these stimuli in the current experiment. I used the same 70 filler items as the previous experiment. These filler items were used to confirm that the two experiments are comparable.

Procedure

The procedure was identical to the previous experiment, except for the instructions. While participants were instructed to articulate a natural word as a continuation as quickly as possible in the previous experiment, participants in this study were instructed to complete each

sentence fragment as quickly as possible. This was because there were more multiple-word expressions in the target verbs Ehrenhofer et al. (2019) used compared to Chow, Smith, et al. (2016), and their offline cloze norming survey asked the participants to complete sentence fragments. I later demonstrate that this slight change in the instructions had minimal impact on the way participants were engaged in the task.

Data Processing

The recordings were processed in the same way as the previous experiment, except that an undergraduate RA corrected the initial transcription and onsets generated by the TransCloze pipeline.

Exclusion

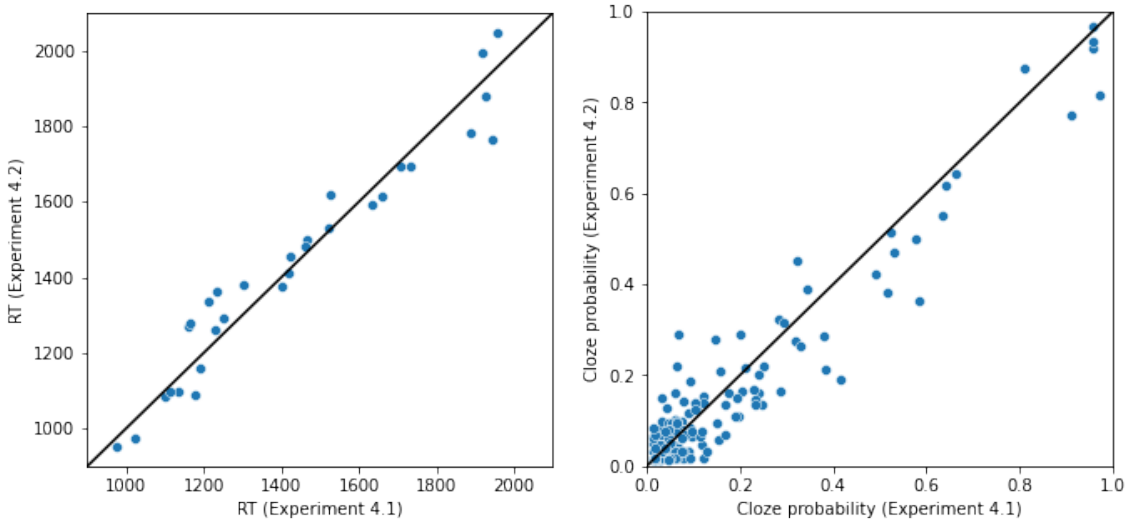
Participants and anomalous trials were excluded in the same way as the previous experiment. This left 64 participants and 3222 responses, which corresponds to 83.9% of all responses from the 65 participants.

4.2.2 Results

Filler items

I first analyzed 30 out of the 70 filler items shared between Experiment 4-1 and 4-2 to confirm that the two experiments are comparable, despite the differences in the instructions. The mean cloze RTs were 1415ms for Experiment 4-1 and 1425.36ms for Experiment 4-2. A linear mixed-effects model analysis of the RTs revealed no significant effect of the stimulus sets ($\beta = 5.03, p = .87$). Next, I averaged the RTs for each filler item for each experiment

and plotted them in Figure 36a. The x-axis shows the mean RTs in Experiment 4-1 for each filler item and the y-axis shows the mean RTs for the same item in Experiment 4-2. When the within-context mean RTs are identical in the two experiments, the dots are on the black lines. As the figure illustrates, there was a strong correlation between the RTs for the same filler items in the two experiments ($r = 0.97$). Similarly, I computed the cloze probability of each response and plotted in Figure 36b. As the figure illustrates, there was also a strong correlation between the cloze probabilities in the two experiments ($r = 0.96$). These suggest that the difference in the instructions did not significantly change the cloze RTs or cloze probabilities.



(a) Correlation of within-context mean RTs **(b)** Correlation of cloze probabilities

Figure 36: Comparison of filler item responses in the two experiments. (a) shows the within-context mean RTs, where each dot represents each context. (b) shows the cloze probability of each continuation and each dot is each continuation, i.e., multiple possible continuations per context. If there are perfect correlations between the responses in the two experiments, all the dots should be on the black lines.

Overall proportions of response categories

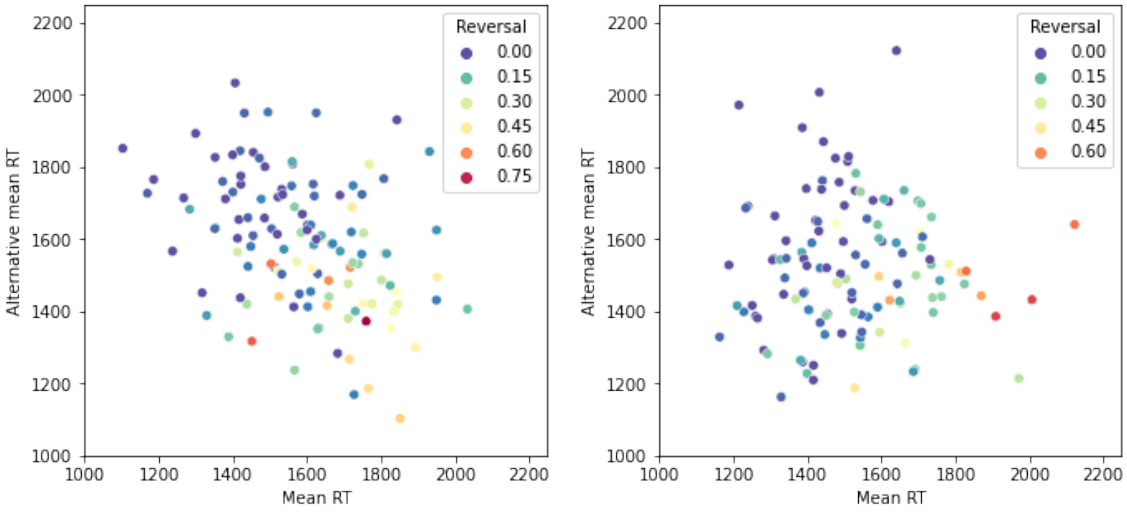
Table 10 summarizes the proportions of the response categories in Experiments 4-1 and 4-2. While there were more role-specific responses in Experiment 4-2, there was the same overall

pattern that (i) role-specific responses were most frequent and role-inappropriate responses were rarest, and that (ii) role-appropriate responses combining role-specific responses and neutral responses accounted for the majority of the responses.

	Specific	Neutral	Inappropriate
Experiment 4-1	59.2%	26.5%	14.3%
Experiment 4-2	65.4%	22.5%	12.1%

Table 10: Proportion of role-specific, neutral and role-inappropriate responses in the two speeded cloze tasks.

The relationship between the mean role-appropriate RTs and the alternative mean role-appropriate RTs is presented in Figure 37b. As can be seen in the figure, there are more contexts with both short role-appropriate RTs and alternative role-appropriate RTs (the bottom left corner) compared to the data in Experiment 4-2 (37a). That is, Ehrenhofer et al. (2019)'s stimuli had contexts with both strong role-appropriate candidates and strong role-inappropriate lures, as they were designed to have.



(a) Experiment 4-1 **(b)** Experiment 4-2

Figure 37: The relationship between mean RT (x-axis), alternative mean RT (y-axis), and the proportion of role-inappropriate responses (color).

Context-wise variability in RTs of role-appropriate responses

Figure 38 shows the relationship between the role-appropriate RTs and the strength of lures estimated using the alternative mean RTs. A linear mixed-effects model analysis conducted in the same way as in Experiment 4-2 revealed a significant effect of the cloze probability ($\beta = -0.52, p < .001$), and the mean alternative RT ($\beta = -0.42, p < .001$). but no interaction between the two effects ($\beta = 0.09, p = .66$). This pattern replicated the findings from Experiment 4-1.

It should be noted that another analysis without the random effects for items showed no significant effect of the mean alternative RT ($\beta = 0.09, p = .083$). This follows the context-wise analyses of Staub et al. (2015), who did not include the random effects for items. Staub et al. made this decision because there were no categorical conditions in their experiment and each item had only one value for context-level variables (modal cloze probability in their analyses). In such cases, the random effects of items cannot be estimated independently of the context-level fixed effects. On the other hand, since each item in our experiment was presented in two different word orders, the random effects of items can be independent of the context-level fixed effects (i.e. mean alternative RT). Therefore it is more reasonable to include the random effects for items in the current experiments. However, the effect of the mean alternative RT, which is a measure of the lure strength, should be taken with caution because whether or not to include the random effects of items resulted in a drastic difference in the fixed effects. Additionally, RT analysis in Experiment 4-1 did not change by this modification in the random structure.

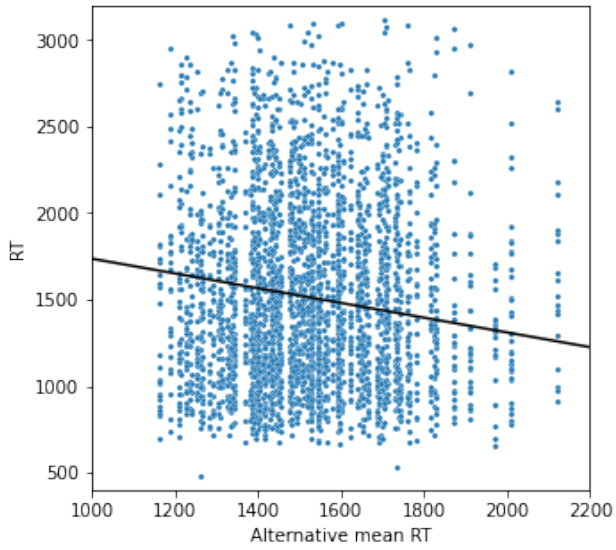


Figure 38: The relationship between the alternative mean RTs (x-axis) and the role-appropriate RTs (y-axis) for Experiment 4-2.

Context-wise variability in the role-inappropriate responses

The proportion of role-inappropriate responses produced for each context and its relationship with the mean role-appropriate RTs and the alternative mean role-appropriate RTs are presented in Figure 37b. A binomial generalized linear mixed-effects model revealed a significant effect of the mean RTs of the context itself ($\beta = 5.81, p < .001$), and the mean RTs of the alternative context ($\beta = -2.62, p < .001$), but no interaction between the two effects ($\beta = 5.28, p = .126$). Figure 39 illustrates the model predictions. The probability of role-appropriate responses increases with longer mean role-appropriate RTs and shorter alternative mean role-appropriate RTs. This suggests that role-appropriate responses are slower in contexts with stronger lures.

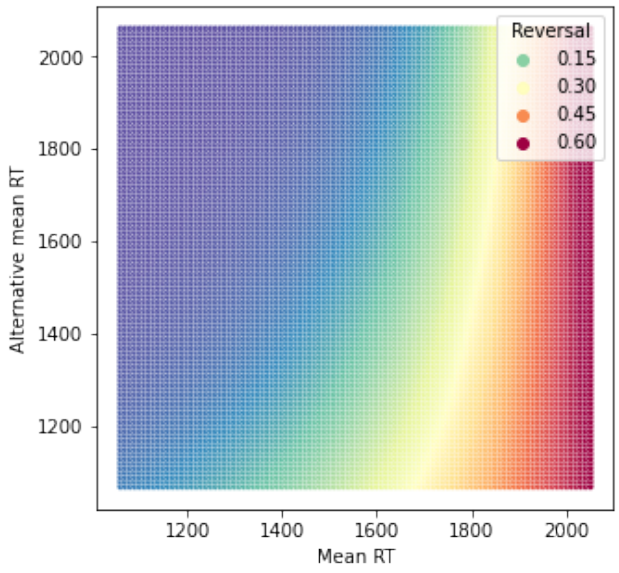


Figure 39: An illustration of the binomial generalized linear mixed-effects model's prediction. The axes and the color are identical to Figure 37b

4.2.3 Discussion

Experiment 4-2 combined with Experiment 4-1 provides evidence for role-inappropriate activation via the between-context variability in lure strengths. The availability of lures in each context was estimated using the mean role-appropriate RTs in the alternative context. Contexts with stronger lures had (i) slower role-appropriate responses and (ii) more role-inappropriate responses. These findings suggest that role-inappropriate candidates are pre-activated such that they can inhibit the role-appropriate candidates and slow down those responses and that they can win races at least occasionally, leading them to be produced.

The analysis of the relationship between the role-appropriate and the alternative mean RTs suggests a novel use of (speeded) cloze data. In cloze tasks, we can only use cloze measures from lexical candidates that are produced by at least one participant. This nature of the task makes it difficult to investigate lexical candidates that are rarely or never

produced by participants, including but not limited to argument role reversals. However, as I demonstrated in the current studies, the pre-activation of such candidates can be indirectly measured via the amount of inhibition they impose on the other candidates that are produced.

The observation that there are more role-inappropriate responses when there stronger lures can be straightforwardly explained if role-inappropriate candidates are pre-activated. However, the data might still be explained by an account in which people do not consider and pre-activate role-inappropriate candidates in general, but where they occasionally produce such candidates when they misinterpret the contexts, if that misinterpretation correlates with the availability of role-appropriate candidates in the alternative contexts. For example, people might misinterpret (25a) as (25b) more often than to misinterpret (26a) as (26b). Although this possibility cannot be excluded by the current experiments, it is unlikely, because people have to misinterpret (25a) more often than (26a) without the verb. In the speeded cloze task, participants are not directly presented with lures in the (speeded) cloze experiments and they do not predict those lures under this hypothesis. Therefore, this account requires that contexts such as (25a) are more confusing than (26a), which does not seem to be likely.

(25) a. The restaurant owner forgot which customer the waitress had ...

b. The restaurant owner forgot which waitress the customer had ...

(26) a. The parent saw which child the lifeguard had ...

b. The parent saw which lifeguard the child had ...

4.3 Implications for the monitoring mechanism

The data collected in the two speeded cloze studies also have implications for the way role-inappropriate responses are produced under an approach that gives an important role to a monitoring mechanism. In the role-insensitive pre-activation hypothesis, role-inappropriate candidates are pre-activated just as strongly as role-appropriate candidates. However, a monitoring mechanism checks for the plausibility of the utterance being produced and prevents role-inappropriate productions.

The occasional role-inappropriate responses in the speeded cloze data can be explained in at least two ways. First, they might reflect failures of detection of role-inappropriate winners. While role-inappropriate winners of races are assumed to be detected by the monitoring mechanism in most cases, it might fail to detect those role-inappropriate winners occasionally. Second, role-inappropriate responses might be produced as a last resort when there are no other candidates to be produced. In the speeded cloze task, there is some temporal pressure on participants to produce cloze responses as quickly as possible. If a role-inappropriate candidate is highly pre-activated but role-appropriate candidates are not, there should be some cases where only role-inappropriate candidates reach the threshold, and where there is pressure to quickly produce a response. In such cases, people might choose to produce role-inappropriate candidates as a last resort in order to provide some response, even if they are aware that the available winners are role-inappropriate.

The two mechanisms in fact make different predictions about the latencies of role-inappropriate responses compared to the latencies of role-appropriate responses. If role-inappropriate candidates are predicted as strongly as role-appropriate candidates, and the role-inappropriate

responses are produced because they somehow escaped the monitoring process, then role-inappropriate responses should be faster than role-appropriate responses. Assuming that role-inappropriate candidates are just as strongly pre-activated as role-appropriate candidates, but role-appropriate responses are produced much more frequently, a large number of the role-appropriate responses should be products of revision processes. That is, they would be produced after a role-inappropriate candidate reaches the threshold but is subsequently inhibited. Therefore, a large number of role-appropriate responses should be produced following such a revision process, resulting in longer RTs on average. In the same scenario, role-inappropriate responses should be produced faster than role-appropriate responses, because they cannot be products of such a revision process. Therefore, role-inappropriate responses on average should be faster than role-appropriate responses.

On the other hand, if role-inappropriate responses are produced as a last resort when there is no other response to produce, the role-inappropriate candidates should be slower than role-appropriate responses. While role-appropriate responses can be produced any time a role-appropriate candidate reaches the threshold and passes the monitoring process, role-inappropriate candidates are produced when there is no such candidate that was able to be produced earlier. Hence role-inappropriate responses should be slower than role-appropriate responses.

The RT data of Experiments 4-1 and 4-2 can be used to evaluate these predictions. The mean cloze RTs of each response category for each experiment are presented in Table 11 and Figure 40. A linear mixed-effects model of cloze RTs revealed a significant effect of role-appropriateness ($\beta = -0.21, p < .001$). There was no significant effect of experiments ($\beta = -0.06, p = .148$) or interaction between experiments and role-appropriateness ($\beta =$

$0.01, p < .695$). This pattern was also seen in other speeded cloze experiments (Lee et al., 2022).

Thus role-inappropriate responses are slower than role-appropriate responses in both experiments. This suggests that role-inappropriate responses were not produced because the monitoring mechanism failed to catch them, but because there were no role-appropriate candidates to be produced.

	Appropriate	Inappropriate
Experiment 4-1	1558	1761
Experiment 4-2	1491	1696

Table 11: Cloze RTs of role-appropriate, neutral and role-inappropriate responses in the cloze tasks.

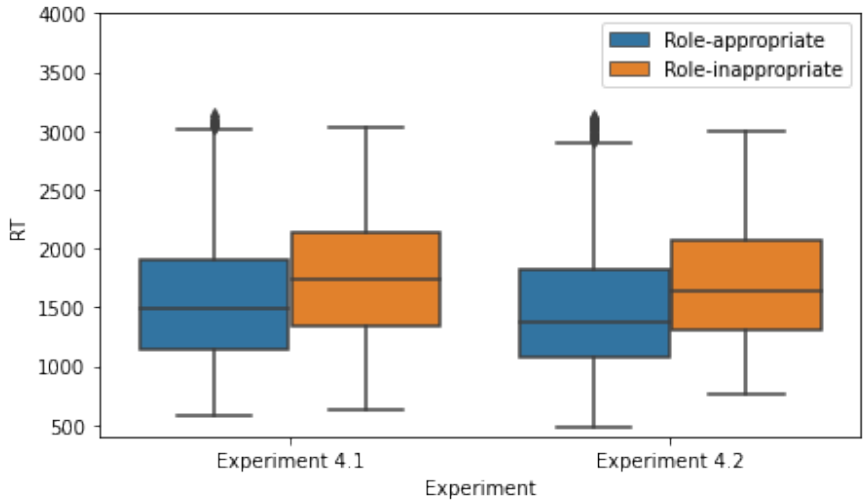


Figure 40: Cloze RTs of role-appropriate and inappropriate responses in Experiments 4-1 and 4-2.

Overall, Experiments 4-1 and 4-2 serve as an initial step in understanding the discrepancy between different measures. Here I asked if there is any evidence for role-inappropriate preactivation, and the results suggest that role-inappropriate candidates are indeed activated sufficiently to inhibit role-appropriate competitors or to win races at least occasionally. Fur-

thermore, within the symmetric pre-activation hypothesis, the comparison of role-appropriate and inappropriate RTs provides evidence for role-inappropriate production as a last resort, rather than as a consequence of monitoring failure.

However, the current studies do not provide evidence for whether the pre-activations of role-inappropriate candidates are as large as those of role-appropriate candidates, or whether they are still much smaller than the activations of role-appropriate candidates. In order to test if there is matched pre-activation for role-appropriate and inappropriate candidates, I turn to two different pre-activation measures in the next chapter.

5 Lexical tasks and distributional analyses of reaction times

Previous chapters have explored the mismatch between measures of pre-activation. Chapter 3 demonstrated that there is a mismatch between N400 amplitudes and speeded cloze probabilities in argument role reversals, even when both stimuli and time courses are carefully matched. While there was no contrast between the N400 amplitudes for role-appropriate and inappropriate candidates when the target verbs were presented shortly after the contexts, speeded cloze responses were predominantly role-appropriate. Therefore, the different measures tell inconsistent stories about whether and how argument roles impact pre-activation. The N400 studies suggest that argument roles do not influence pre-activation at least temporarily and hence that role-appropriate and inappropriate candidates are matched in terms of pre-activation. On the other hand, the speeded cloze study suggests that argument roles have a large impact on pre-activation, and role-appropriate candidates are much more pre-activated compared to role-inappropriate candidates. Chapter 4 showed that there is some degree of role-insensitive pre-activation in the speeded cloze task, using context-wise variability in the estimated availability of role-inappropriate candidates ("lures"). Role-appropriate responses with stronger lures were produced more slowly, presumably reflecting lateral inhibition from the pre-activated lures. Additionally, there were more role-inappropriate responses with stronger lures, suggesting that role-inappropriate candidates can at least occasionally win races and be produced in the speeded cloze task.

While the previous studies in this thesis provided evidence for some amount of role-inappropriate pre-activation, it is not clear how much role-inappropriate candidates are pre-

activated compared to role-appropriate candidates. They might be pre-activated just as much as role-inappropriate candidates, suggesting that argument roles have no impact or minimal impact on pre-activations. On the other hand, they might be much less pre-activated than role-appropriate candidates, where argument roles can have a strong influence on pre-activations. In this chapter, I seek more direct evidence of role-(in)sensitive pre-activation, in order to evaluate the influence of argument roles on pre-activation. While N400 amplitudes and speeded cloze responses might be influenced by post-lexical combinatorial processes, I use other measures of pre-activation that are potentially less susceptible to post-lexical processes, and I further isolate the effects of pre-activation by analyzing the distributions of reaction times.

5.1 Lexical decision task and Read-aloud task

If the conflicting pre-activation measures rather transparently reflect the common underlying pre-activation, as assumed in the Simple Activation Model (Chapter 1), those measures should align with each other in any case, including argument role reversals. In order to account for the mismatch between the pre-activation measures, I proposed two hypotheses that question the linking hypotheses in different ways. The symmetric pre-activation hypothesis maintains that cloze responses do not simply reflect pre-activations, and that cloze responses are skewed by a post-lexical combinatorial process. I proposed in Chapter 3 that once a lexical candidate is pre-activated and reaches a threshold, it is temporarily incorporated into the utterance meaning representations and a monitoring mechanism checks for the plausibility of the lexical candidate in its context. With such processes, role-inappropriate candidates are detected and inhibited before they are produced even if pre-activations are matched

across role-appropriate and inappropriate candidates. Thus the discrepancy between the pre-activation measures might be caused by the monitoring mechanism. On the other hand, under the asymmetric pre-activation hypothesis, the linking hypothesis of N400 amplitudes is instead questioned. Some researchers argue that it reflects post-lexical combinatorial processes. Some of them maintain that N400 amplitudes reflect the processes to integrate the meanings of lexical items into the utterance meaning representations (e.g. Brown & Hagoort, 1993; Rabovsky et al., 2018), while others argue that N400 amplitudes reflect the detection of semantic anomalies in the utterance meaning representations (e.g. Kim & Osterhout, 2005). If N400 amplitudes in fact reflect these post-lexical processes, they do not align with cloze probabilities because they reflect different processes.

Given this uncertainty about the linking hypotheses for the available measures, if there are alternative measures that can assess the pre-activation without the influence of post-lexical combinatorial processes, they would provide better evidence for role-(in)sensitivity in pre-activation processes. I propose that simple lexical decision and read-aloud (naming) tasks may be more transparent measures of pre-activation, excluding the influence of post-lexical processes. In the lexical decision task, participants are asked to judge if a presented string of letters is a real word in the target language or is a meaningless non-word. Response latencies in this task are faster when the lexical representations are pre-activated, for example by semantic priming (Meyer & Schvaneveldt, 1971; Perea & Rosa, 2002; Ratcliff et al., 2004). It is also known that sentential contexts can facilitate lexical decisions, and that RTs in the lexical decision task are faster for high-cloze items than low-cloze items (Kleiman, 1980). In the read-aloud task (or the naming task or the pronunciation task), people simply read aloud a presented string of letters, and the latency of the speech production onset is used as

the measure. Similar to the lexical decision task, more lexical activation enables articulation with shorter latencies (e.g. Stanovich & West, 1983). RTs in the read-aloud task are also affected by sentential contexts (Seidenberg et al., 1984).

These tasks are different from other measures in that responses can be initiated once basic lexical processes are completed. In the lexical decision task, participants only need to access the lexical representation for the target lexical item and find out that it is in the lexicon. The read-aloud task can be completed once the lexical representation has been accessed and the form associated with it has been retrieved. Therefore, the nature of these tasks might reduce the influence of post-lexical processes.

Importantly, nothing about the task demands in these studies requires the monitoring mechanism to quickly inhibit role-inappropriate candidates, unlike the speeded cloze task. As discussed in Chapter 3, the monitoring mechanism might be involved in tasks other than the speeded cloze task, but the threshold for monitoring is lowered in the speeded cloze task. In the speeded cloze task, people are not presented with continuations, and they themselves must provide responses quickly. Under such circumstances, if no lexical item reaches the threshold quickly, people would not be able to produce anything. A lowered threshold can increase the chance of successful production under time pressure. It simultaneously allows quick detection and inhibition of role-inappropriate candidates by the monitoring mechanism, which is triggered by candidates reaching the activation threshold. On the other hand, in EEG studies with comprehension tasks, as well as the lexical decision task and the read-aloud task, the threshold could be high because people do not provide responses by themselves. They might pre-activate lexical candidates, but there is no pressure to have lexical candidates reach the threshold, which results in slower inhibition of role-inappropriate candidates.

However, the lexical decision task and the read-aloud task might not completely block the influence of post-lexical combinatorial processes, even though they do not require those processes. Participants can, in principle, incorporate the meaning of the presented lexical item into an utterance meaning representation before responding to the lexical decision or read-aloud tasks, even if they are not required to do so. If this is the case, the effect of pre-activation on RTs might be obscured by the effect of combinatorial processes. In order to isolate the effect of pre-activation on RTs from the combinatorial processes, I employ distributional analyses of the RTs in the two tasks, as described in detail in Section 5.2.

5.1.1 A prior study in the read-aloud task

There is one previous study that used the read-aloud task for argument role reversals. Ferretti, McRae, and Hatherell (2001) conducted a cross-modal read-aloud task in their Experiment 4. They presented participants with contexts auditorily, such as the parts preceding the slashes in (27), (28) (29), and (30). The target words (*crook* or *cop* in the examples) were presented visually at the offset of the contexts and participants read them aloud. (27b) and (29b) can be considered as argument role reversal sentences in that *cop* is assigned an inappropriate role (patient), although it would be an appropriate agent. (28) and (30) can be considered as baselines for (27) and (29) respectively. Ferretti et al. found that target words that are appropriate in the contexts ((27a) and (29a)) were produced faster compared to the baselines, but the same target words in role-inappropriate contexts ((27b) and (29b)) were not. This suggests that role-appropriate continuations such as *crook* for (27a) or *cop* for (29a) were pre-activated using the contexts, leading to shortened RTs, but that role-inappropriate continuations such as *cop* for (27b) or *crook* for (29b) were not. This seems to be inconsistent

with the findings in EEG studies that repeatedly found no contrast between role-appropriate and inappropriate continuations in role reversal contexts, including the Japanese study by Momma (2016) re-analyzed in Chapter 3 and other prior studies (Kolk et al., 2003; Kuperberg et al., 2003; Hoeks et al., 2004; Kim & Osterhout, 2005; Oishi & Tsutomu, 2010; Chow, Smith, et al., 2016; Chow et al., 2018; Liao et al., 2022).

(27) a. She arrested the / crook.

b. She arrested the / cop.

(28) a. She kissed the / crook.

b. She kissed the / cop.

(29) a. She was arrested by the / cop.

b. She was arrested by the / crook.

(30) a. She was kissed by the / cop.

b. She was kissed the / crook.

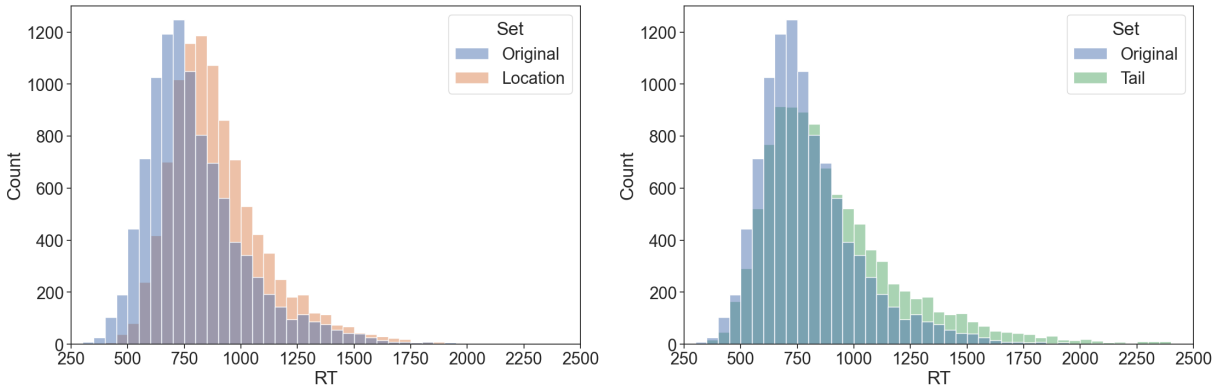
A few differences between Ferretti et al. (2001)'s study and other prior studies on argument role reversals should be pointed out. First, while most other prior studies on argument role reversals investigated the processing of verbs given arguments, arguments were the targets (*cop* or *thief* in the examples above) and verbs were provided to participants in this study. People could have been in different states after encountering a context that included a verb (e.g., Friederici & Frisch, 2000). Second, it is not clear how much the time courses were matched with other prior studies. Since the stimuli were presented auditorily, the timing was not controlled in the same way as in prior EEG studies where the stimuli were presented in

rapid serial visual presentation. The time course of the experiment is critical for interpreting the results, given the findings about effects of timing on N400 effects (Chapter 3 and Chow et al. (2018); Liao et al. (2022)) as well as in the visual world paradigm (Kukona et al., 2011). Given that participants were presented with the target arguments at the offset of the context, they did not have a long interval between the presentation of the context and the target argument, but it is not clear how much the time courses aligned with other studies. A study that is better matched with prior studies will provide better evidence for or against role-sensitive pre-activation. Furthermore, as mentioned above, the read-aloud task may not completely exclude the involvement of post-lexical combinatorial processes. Distributional analyses of RTs will allow me to investigate the influence of pre-activation on RTs independent of the combinatorial processes.

5.2 Distributional analyses of reaction times

In order to isolate the potential influence of pre-activation and of post-lexical combinatorial processes, I conducted response time distributional analyses (e.g. Ratcliff, 1979; Balota, Yap, Cortese, & Watson, 2008; Staub, 2011). The analyses built on the fact that differences in right-skewed distributions such as RTs can be caused by differences in the locations of the entire distributions (Figure 41a), or by differences in the tails of the distributions (Figure 41b). These two types of differences between RT distributions can be dissociated by vincentile plots or by fitting distributions such as ex-Gaussian distributions to the data. I propose that pre-activation should mainly modify the locations of the RT distributions, while post-lexical combinatorial processes such as monitoring or semantic anomaly detection should primarily influence the tail of the distribution.

When researchers compare RTs in two conditions, they often calculate the means of the distributions and compare them between the conditions. However, RT distributions are very rarely symmetric. They are usually right-skewed (e.g., distributions in Figure 41). A difference between the means of two right-skewed RT distributions can be caused by a shift in the location of the distribution and/or a larger and extended tail. For example, the orange and green distributions are generated by shifting the location of the blue distribution (Figure 41a) and by modifying the tail of the blue distribution (Figure 41b). Despite the difference in the shapes, the orange and green distributions have identical means.



(a) Shifted location **(b)** Increased tail

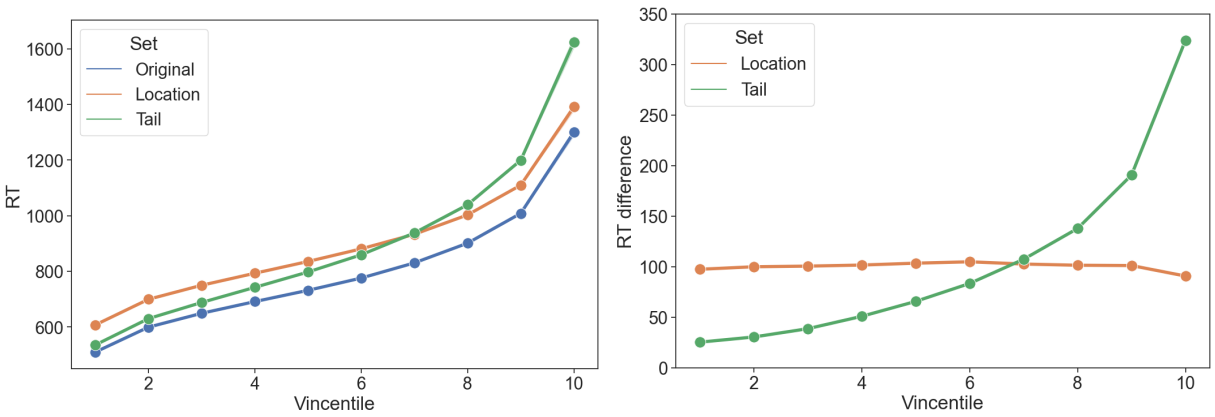
Figure 41: Differences in RT means can be caused (a) by shifting the location or (b) by modifying the tail. The blue distributions in the two figures are identical and the orange and green distributions have matched means.

Some researchers focused on this distinction between the location and the tail and investigated whether a certain factor influences the location or the tail of the RT distributions (e.g. Balota et al., 2008; Staub, 2011). That is, whether the factor influences essentially all responses and shifts the entire distribution, or whether it strongly influences a subset of responses and right-skews the distribution by modifying the tail.

I propose that pre-activation should influence the location, while integrative processes

should influence the tail in the lexical tasks. This is because increased pre-activation for the target lexical item should facilitate essentially all responses for that item. On the other hand, presumably, the integrative processes in the lexical tasks are not deployed in most cases because they are not required by the task, and therefore they should more strongly influence only a subset of responses. In fact, Balota et al. (2008) showed that semantic priming shifts the RT distributions in both lexical decision and read-aloud tasks. Thus, if role-inappropriate candidates are pre-activated just as much as role-appropriate candidates, the locations of the distributions should be closely matched, while the tails might or might not be different. On the other hand, if the two classes of candidates are pre-activated differently, the locations of the RT distributions should reflect that contrast.

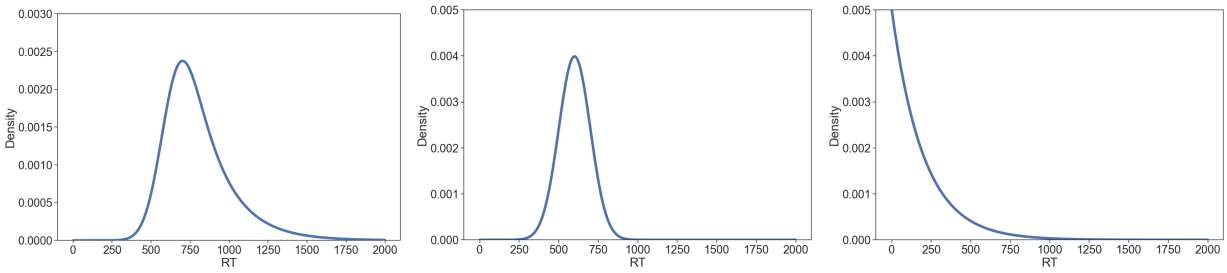
Prior studies have used two ways to isolate the effects on the location and the tail. One approach is to use vincentiles. This approach ranks the RTs and divides the data into bins of equal numbers of observations based on the ranks. For example, the shortest 10% of RTs (per condition per participant) are in the first bin, the next 10% are in the second bin, and so on. Vincentile refers to the mean of the RTs in each bin. Figure 42a shows the vincentile plots for the same data sets used in Figure 41. Figure 42b shows the difference between the vincentiles of the original distribution (blue) and two modified distributions (orange and green). A shift in the location of the RT distributions can be observed as a difference that is consistent across bins (Figure 42a: blue vs orange). In the difference plot, a pure contrast in locations appears as a horizontal line (Figure 42b: orange). On the other hand, the larger and extended tail can be observed as an increased slope and an especially large increase in the slowest bins (Figure 42a: blue vs green). In the difference plot, this appears as no difference in the first bins but increasingly greater differences in the later bins (Figure 42b: green).



(a) Vincentile plots **(b)** Difference plots

Figure 42: Vincentile plots of the simulated data in Figure 43a. (b) shows the differences between the vincentiles of the original distribution (blue dots/lines in (a)) and the vincentiles of the modified distributions (orange and green dots/lines in (a)).

Another approach is to fit distributions such as ex-Gaussian (exponentially modified Gaussian) distributions to the RT distributions (Ratcliff, 1979). Ex-Gaussian distributions are right-skewed distributions that are convolutions of two distributions. One of the components is a Gaussian distribution (normal distribution) and the other is an exponential distribution. For example, an ex-Gaussian distribution in Figure 43a can be considered as a convolution of the two distributions illustrated in Figures 43b and 43c. There are three parameters for the two components: μ and σ for the Gaussian component, which respectively captures the mean and the standard deviation of the component, and τ for the exponential component, which controls both the mean and standard deviation of the component. A shift in the location of the distribution can be captured as a difference in the μ parameter, while a difference in the right tail can be captured as a difference in the τ parameter. Therefore, the effect of a factor on two right-skewed RT distributions can be isolated by fitting ex-Gaussian distributions to the RT distributions and comparing the μ and τ parameters of the two ex-Gaussian distributions.



(a) Ex-Gaussian **(b)** Gaussian component **(c)** Exponential component

Figure 43: An ex-Gaussian distribution (a) consists of the Gaussian component (b) and the exponential component (c).

In this study, I conduct vincentile analyses of the RTs in the lexical decision task and the read-aloud task. I did not conduct ex-Gaussian analyses because the limitations in the experiment designs did not provide sufficient observations per participant or per item per condition to fit ex-Gaussian distributions to RT data. (Ratcliff (1979) recommends data with 100 or more observations to fit a gamma distribution for reliable fitting). Therefore, I only report the vincentile analyses that are suitable for analyses with fewer observations (Ratcliff, 1979).

5.3 Experiment 5-1

The current study tested for the pre-activation of role-inappropriate verbs in the lexical decision task and the read-aloud task. In order to match the stimuli and the time course with prior studies on argument role reversals, I used the same stimuli based on Chow, Smith, et al. (2016) and presented them in the same manner as Chow, Smith, et al.. Participants either judged whether the target verb is a real English word or a non-word in the lexical decision task, or simply read them aloud in the read-aloud task.

The stimuli consisted of two types of sentence pairs: argument role reversal items (reversal

items: e.g., (31)) and argument substitution items (substitution items: e.g., (32)). The reversal items had the same arguments but with different argument role assignments. On the other hand, substitution items had different direct object arguments. The target verbs, *served* and *evicted* in the examples, had high offline cloze probabilities in the high-cloze condition ((31a) and (32a)). On the other hand, the same verbs had low offline cloze probabilities in the low-cloze condition ((31b) and (32b)). The offline cloze probabilities were matched between the two high-cloze conditions and between the two low-cloze conditions.

(31) a. The restaurant owner forgot which customer the waitress had served.

b. The restaurant owner forgot which waitress the customer had served.

(32) a. The superintendent overheard which tenant the landlord had evicted.

b. The superintendent overheard which realtor the landlord had evicted.

The primary interest is in the comparison between the reversal/high-cloze condition (31a) and the reversal/low-cloze condition (31b). If pre-activations are matched for role-appropriate and inappropriate verbs, the locations of the RT distributions should be matched for (31) and any differences in the mean RTs should arise from difference in the tails, if at all. On the other hand, if role-appropriate verbs are more strongly pre-activated than role-inappropriate verbs, there should be a contrast in the locations of the distributions. The substitution items (32) serve as baselines in which participants should pre-activate the target verbs differently in the high-cloze and low-cloze conditions. Chow, Smith, et al. in fact found N400 contrasts between the substitution items, while they did not find N400 contrasts between the reversal items. Therefore, there should be a contrast in the locations of the RT distributions of the substitution items.

5.3.1 Methods

Participants

80 native American English speakers were recruited online on the crowdsourcing platform Amazon Mechanical Turk. Participants received monetary compensation. Four participants' data were not analyzed due to missing data. Additionally, five participants were excluded from the data due to unsatisfactory performance as described below. This left 71 participants for analysis (42 females and 29 males; average age = 40.9). Another 2 participants were excluded from the analyses of the read-aloud task due to poor audio quality.

Materials

120 sentence pairs of high-cloze and low-cloze sentences were taken from Chow, Smith, et al. (2016)'s Experiment 1. The stimuli in the current experiment were identical to the original stimuli up to the target verb. However, the stimuli in this experiment did not include the parts following the target verbs so that they do not interfere with the lexical decision task or the read-aloud task. 60 pairs of stimuli were reversal items such as (31) and the other 60 pairs were substitution items such as (32). The low-cloze condition of the reversal items was created by reversing the arguments of the high-cloze condition (31b), while the low-cloze condition of the substitution items was created by replacing the object of the high-cloze condition (32b). One of the pairs from Chow, Smith, et al. was replaced with another sentence pair due to potentially sensitive content. For each participant, 30 reversal items and 30 substitution items were presented in the lexical decision task and the remaining stimuli were presented in the read-aloud task. Half of the 30 stimuli sets were presented in the high-cloze condition

and the rest were presented in the low-cloze condition. Therefore, there were 15 experimental items for each condition per task per participant.

Additionally, 60 filler items were created for the read-aloud task. Half of the fillers were plausible and the others were implausible. The target words to be read aloud were always the final words of the sentences. Another set of 180 filler items was created for the lexical decision task. 60 of those fillers had embedded object questions and had OSV word order (henceforth OSV fillers) and the targets were non-words. Since all the experimental items had embedded object questions ((31) and (32)) and all had real-word targets, the OSV fillers were necessary to make it impossible for participants to expect real words following embedded object questions. 60 of the other fillers were sentences without embedded object questions and with real-word targets. Half of those fillers were plausible and the other half were implausible. Another 60 fillers were without embedded object questions but with non-word targets.

The target non-words were pronounceable pseudo-words either directly obtained from the ARC Nonword Database (Rastle, Harrington, & Coltheart, 2002) or created by combining two pseudo-words in the database. The target non-words in the OSV fillers were created by adding past-participle suffixes to the non-words (e.g. *donsed*, *spriswalved*). This made it impossible for participants to make lexical decisions by attending to the right edge of the target verbs or non-words. The number of letters of the non-words in the OSV fillers was matched with the length of the target verbs of the experimental items. The target non-words in other fillers were created in a similar way, but they appeared with inflectional suffixes if the contexts required them. The numbers of letters were matched with the target words of the filler items with the real-word targets. Finally, a native speaker of American English checked all the non-word stimuli and confirmed that they were in fact non-words.

Procedure

The experiment was conducted via the online experiment platform PCIbex (Zehr & Schwarz, 2018). Stimuli were presented in an almost identical manner as Experiments 4-1 and 4-2. In each trial, a fixation cross was presented on the center of the screen for 800ms. Each word of the sentence fragment appeared word-by-word in black font color on the center of the screen. Each word was presented for 300ms and was followed by a blank screen presented for 230ms, matching the time parameters in Chow, Smith, et al. (2016). Therefore the stimulus onset asynchrony (SOA) was 530ms.

The experiment consisted of two blocks corresponding to the two tasks. In the lexical decision block, the target word in each trial was presented with blue font color, and participants judged if the target word was a real word in English or not by pressing keys. In the read-aloud block, the target word in each trial was presented with red font color, and participants read the target word aloud while their responses were recorded. The audio was recorded for 3000ms following the presentation of the final word, and then a text indicating the end of the trial was presented. The responses were recorded by PCIbex and were automatically sent to a server for data collection. The order of the blocks was counterbalanced across participants.

Some lexical decision and read-aloud trials were followed by a comprehension task where participants were asked about the plausibility of the sentences. They pressed keys and made binary judgments about whether the sentence was plausible or not. This task was necessary to confirm that participants are paying attention to the contexts because, in principle, it is possible to perform the lexical decision and read-aloud tasks without reading the contexts at

all. The comprehension questions followed all the experimental trials, half of the filler items in the read-aloud task and half of the filler items with real words in the lexical decision task. This accounted for 37.5% of all the lexical decision trials and 75% of the read-aloud trials.

Data Processing

For the read-aloud task, the onsets of the speech were first automatically detected by Chronset (Roux et al., 2017) using the TransCloze pipeline developed by the author. Then the onsets were manually corrected by the author using Praat (Boersma & Weenink, 2023), through both listening to the audio and visually inspecting the spectrograms. Trials without valid responses were marked for rejection in this process.

Exclusion

All lexical decision responses with RTs shorter than 200ms or longer than 3000ms were removed. These affected 2.6% of all the experimental trials for the 71 participants who were analyzed. Subsequently, responses with inaccurate responses were removed for the RT analyses. These left 3959 lexical decision trials for experimental items. In the read-aloud task, all trials without responses or with non-target responses were excluded. These trials consisted of 0.3% of all the responses by the 69 participants who were analyzed. Subsequently, responses with RTs longer than 1500ms were removed. This affected 1.0% of all the experimental trials. These processes left 4084 read-aloud trials for experimental items.

Data analysis

The distributional analyses were conducted in different ways from prior studies (Balota et al.,

2008; Staub, 2011). Balota and colleagues obtained vincentiles for each participant for each condition and they analyzed the distributions of the by-participant vincentiles. In the current study, I plotted by-item average RTs instead of by-participant average RTs in the vincentile plots. This was because some participants had as few as four observations for one condition in a task. On the other hand, there were at least twelve observations for each context for each condition and task. There were 16.5 observations in the lexical decision task and 17.0 observations in the read-aloud task for each context for each condition on average. Since the number of observations was still relatively small, I divided responses for each condition for each item into five bins (c.f. Staub (2011) had 38.7 responses on average per condition per participant and had ten bins).

Linear mixed-effects model analyses were conducted using the lmer function of the lme4 package (Bates et al., 2015) and lmerTest package (Kuznetsova et al., 2017) of the statistical analysis software R (R Core Team, 2020). For statistical analyses for vincentiles, I constructed linear mixed-effects models for the RT by trials and the bins were included as one of the fixed factors. In all the analyses below, categorical factors were coded by -0.5 or 0.5 (Substitution: -0.5, Reversal: 0.5; high-cloze: -0.5, low-cloze: 0.5). Bins in the distributional analyses were coded from 0 to 1 with increments of 0.25 in ascending order. That is, the first bin was coded as 0, the second bin was coded as 0.25, and the fifth bin was coded as 1. I included the maximal random effects structure for subjects and items that converged. Under this coding system, the effects on the locations of RTs can be seen as terms without a bin (i.e., fixed effects of cloze manipulation, sentence type, and their interaction). On the other hand, any tail effects should be visible as interactions including bin. All the statistical models for distributional analyses revealed significant effects of bins. That is, responses in faster bins

were faster. I do not discuss these in the Results section because they are simple consequences of how the bins are defined.

5.3.2 Results

Comprehension questions

As mentioned in the Participants section, five participants answered the comprehension questions of the filler items with 66.7% or lower accuracy in either task. These participants were excluded from subsequent analyses. The accuracy of the comprehension questions of the remaining participants was 89.9% in the lexical decision task and 86.8% in the read-aloud task.

The accuracy of the comprehension questions in the experimental items was 80.9% for the reversal condition and 78.1% for the substitution condition in the lexical decision task, and 80.0% for the reversal condition and 79.2% for the substitution condition in the read-aloud task. The rates of plausible judgments in each task in each condition are summarized in Table 12. There seems to be a bias towards plausible judgments in either task and in either sentence type (reversal or substitution). However, people seem to be accurately responding to the comprehension question on average.

Task	**Lexical decision**		**Read-aloud**	
Condition	High-cloze	Low-cloze	High-cloze	Low-cloze
Reversal	93.7%	31.8%	91.8%	31.8%
Substitution	93.9%	37.7%	93.5%	35.2%

Table 12: The accuracy of comprehension questions in the lexical decision task and the read-aloud task.

Lexical decision task

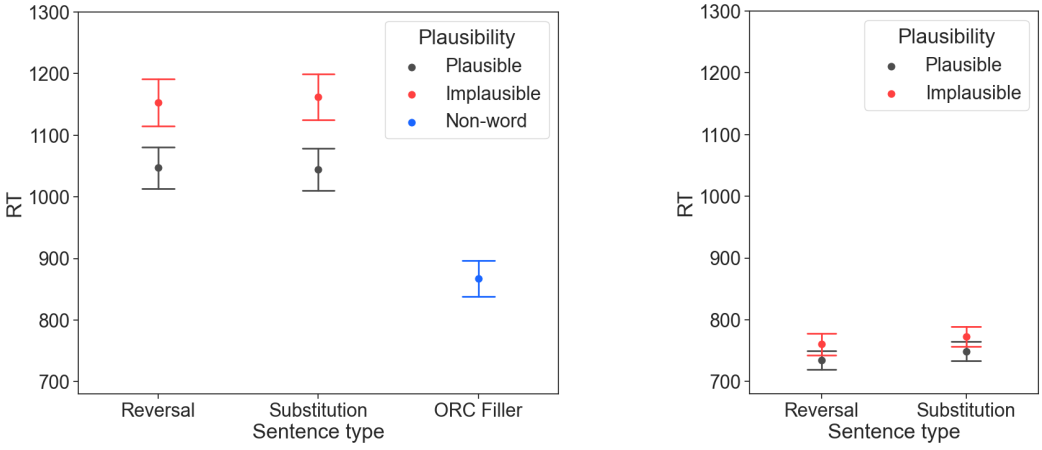
The accuracy of the lexical decision trials is summarized in Table 13. Participants responded accurately to the lexical decision task.

Cloze	High-cloze	Low-cloze
Reversal	99.7%	98.7%
Substitution	99.3%	99.0%

Table 13: Accuracy of the lexical decision task in Experiment 5-1

Non-distributional anlaysis

Inaccurate responses and responses with RTs longer than 2000ms were removed from the non-distributional RT analyses. This process left 3584 trials. The RTs of the accurate lexical responses are shown in Figure 44a. A linear mixed-effects model of RTs revealed a significant effect of cloze manipulation ($\beta = 109.71, p < .001$) but no significant effect of sentence type ($\beta = -4.88, p = .72$) and no interaction between the two effects ($\beta = -16.82, p = .43$).

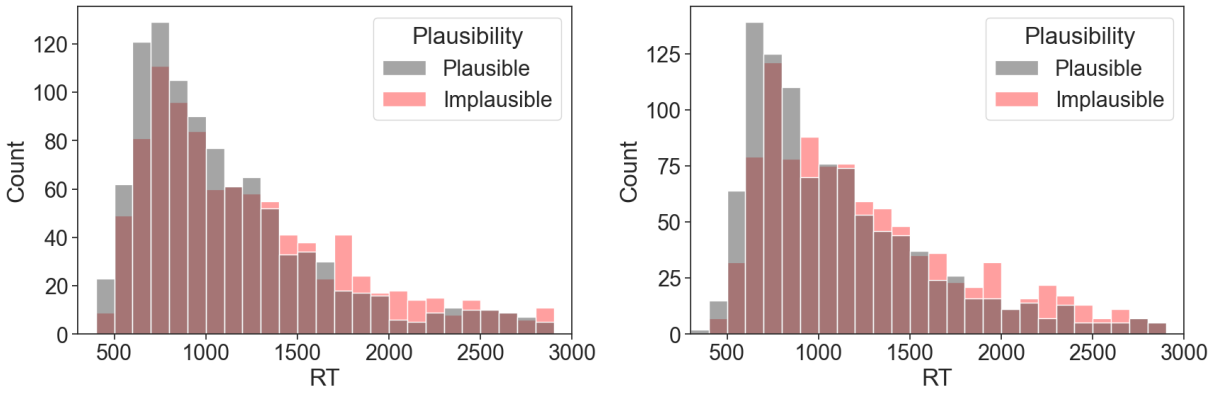


(a) Lexical decision task **(b)** Read-aloud task

Figure 44: RTs of the experimental items (reversal and substitution) in (a) the lexical decision task and (b) the read-aloud task. The RTs of the OSV fillers are plotted for the lexical decision task as well. Error bars represent by-participants standard errors.

Distributional anlaysis

The RT distributions of each condition and each sentence type in the lexical decision task are plotted in Figure 45. The distributions are strongly right-skewed, as we would expect in a task of this kind.

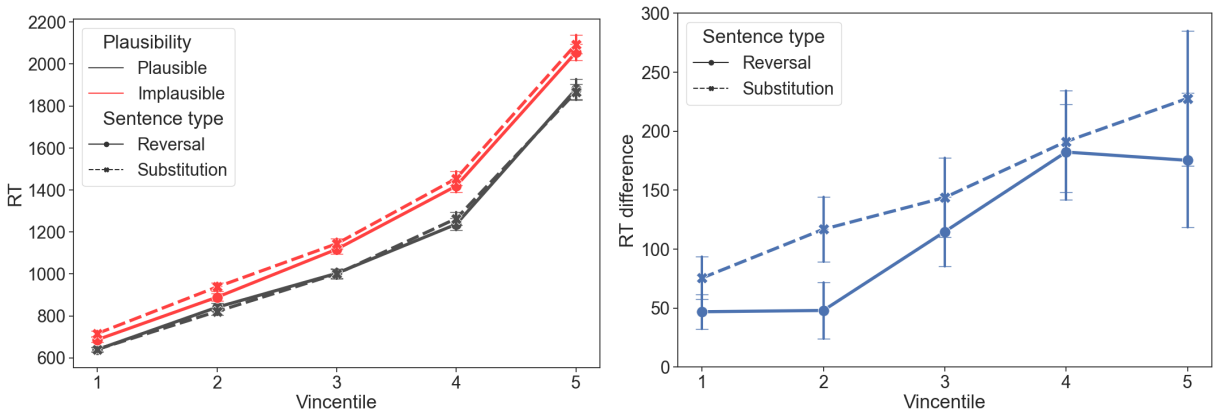


(a) Reversal **(b)** Substitution

Figure 45: The RT distributions of the experimental items with accurate responses in the lexical decision task.

The vincentiles of the lexical decision RTs are plotted in Figure 46a. The differences between the high-cloze condition and the low-cloze condition of each sentence type are plotted in Figure 46b. There appear to be consistent differences between the high-cloze and low-cloze conditions across all bins, as well as slope differences.

I conducted a statistical analysis using a linear mixed-effects model of RTs including cloze manipulation, sentence type, and bin, and all their interactions as predictors. The analysis revealed a significant effect of cloze ($\beta = 54.67, p < .001$), bin ($\beta = 1185.39, p < .001$), and an interaction between the two factors ($\beta = 162.12, p < .001$). There was no interaction between other factors and interactions. These suggest that the cloze manipulation influenced both the locations and the tails of the RT distributions. There was no evidence for contrasts between reversal and substitution stimuli.



(a) Vincentile plots **(b)** Difference plots

Figure 46: Vincentile plots of the lexical decision task in Experiment 5-1. (b) shows the differences between the vincentiles of the original distribution (blue dots/lines in (a)) and the vincentiles of the modified distributions (orange and green dots/lines in (a)).

Read-aloud task

Non-distributional anlaysis

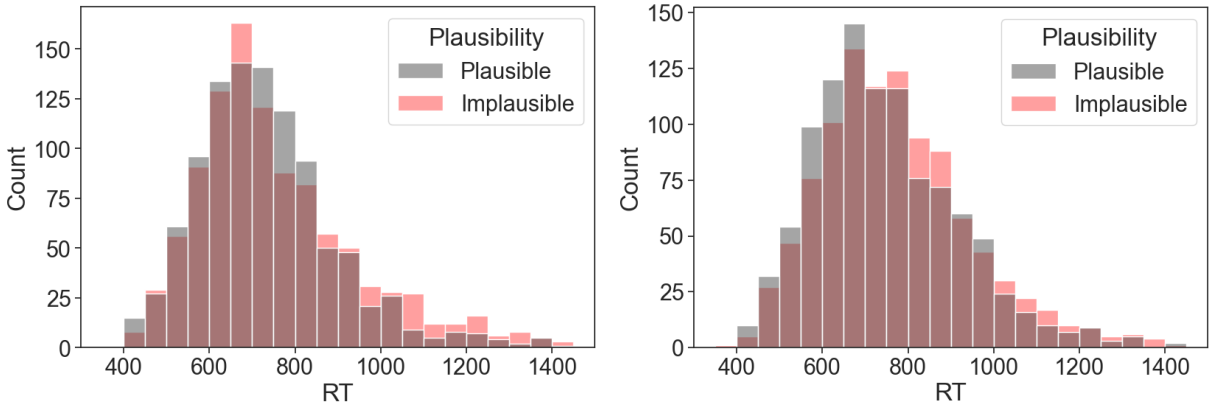
The mean read-aloud RTs are shown in Figure 44b. A linear mixed-effects model of RTs revealed a significant effect of cloze ($\beta = 24.88, p < .001$) but no significant effect of sentence type ($\beta = -13.61, p = .06$) or interaction between the two effects ($\beta = 1.44, p = .84$).

Distributional anlaysis

The RT distributions of each condition and each sentence type in the read-aloud task are plotted in Figure 45. Responses with RTs longer than 1500ms are not included in the figure. The distributions appear to be right-skewed but less so than the lexical decision task.

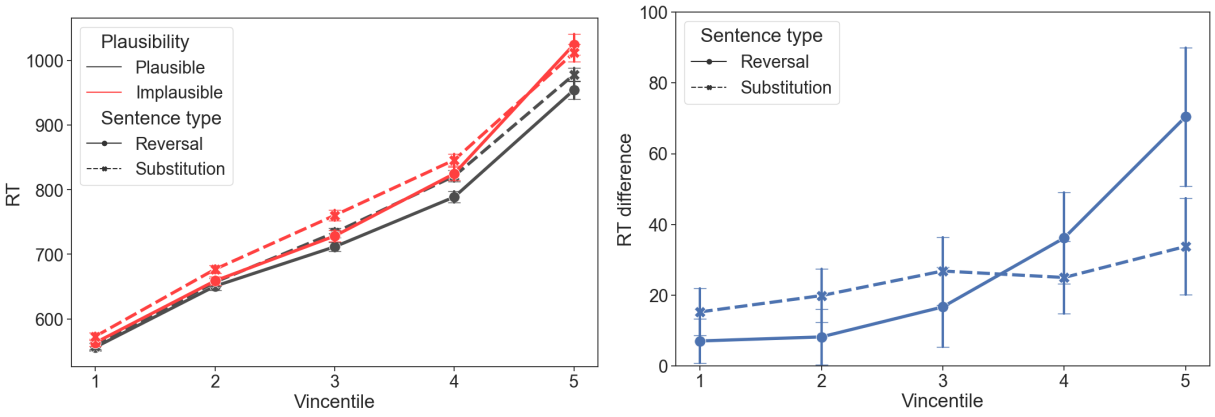
The vincentiles of the read-aloud RTs are plotted in Figure 48a. The difference between the high-cloze condition and the low-cloze condition of each sentence type is plotted in Figure 48b. While there seem to be consistent differences between the high-cloze and low-cloze

conditions of the substitution items across all bins, the difference between the vincentiles of the reversal items seems to be in the slopes.



(a) Reversal **(b)** Substitution

Figure 47: RT distributions of the experimental items with accurate responses in the read-aloud task in Experiment 5-1.



(a) Vincentile plots **(b)** Difference plots

Figure 48: Vincentile plots of the read-aloud task in Experiment 5-1. (b) shows the differences between the vincentiles of the high-cloze and low-cloze conditions (high-cloze RTs subtracted from low-cloze RTs). The solid line represents the RT differences in the reversal items and the dashed line represents the differences in the substitution items.

I conducted a statistical analysis in the same way as for the lexical decision task. The analysis revealed a significant effect of bin ($\beta = 373.58, p < .001$). There was also a significant interaction between cloze and sentence type ($\beta = -19.44, p = .0045$), a significant interaction between condition and bin ($\beta = 40.38, p < .001$), and a three-way interaction between cloze,

sentence type, and bin ($\beta = 43.615, p = .032$). I further fitted models separately for reversal items and substitution items to interpret the interaction between predictability and sentence type. While cloze had a significant effect in the substitution items ($\beta = 16.04, p = .0022$), it did not significantly affect speech latencies in the reversal items ($\beta = -4.48, p = .37$). These analyses suggest that the locations of the RT distribution were different between the substitution/high-cloze condition and the substitution/low-cloze (shorter RTs for the high-cloze condition), but they might not be different between the reversal/high-cloze condition and the reversal/low-cloze condition. These findings were consistent with the pattern in which the cloze manipulation shifts the RT distributions of the substitution items, but does not shift the RT distributions of the reversal items.

The interaction between cloze and bin suggests that the effect of cloze manipulation across the sentence types increased in the later bins. That is, cloze manipulation affected the tails of the distributions. Furthermore, the three-way interaction suggests that this increased effect of cloze manipulation was even larger for the reversal items than the substitution items, as Figure 48b shows.

5.3.3 Disucssion

In order to assess the impact of argument roles on pre-activation processes, the current study conducted the lexical decision task and the read-aloud task and assessed pre-activations for role-appropriate and inappropriate lexical items by distributional analyses of RTs. Pre-activation contrasts should primarily shift RT distributions, which can be seen as constant differences across bins in vincentile analyses. This can be isolated from effects on the tails of RT distributions, which increasingly affect later bins.

The distributional analyses of RTs showed different patterns between the two experiments. Vincentile plots in Figure 46 and the statistical analyses for the lexical decision task showed that the locations of the RT distributions were different between the high-cloze conditions and the low-cloze conditions, regardless of the sentence type (faster RTs for the high-cloze conditions). This suggests that the target verbs in the high-cloze condition were more pre-activated than in the low-cloze condition, not only for the substitution items (e.g., (34)) but also for the reversal items (e.g., (33)). This pattern is compatible with the symmetric pre-activation hypothesis, where the role-appropriate and inappropriate items have matched pre-activation.

(33) a. The restaurant owner forgot which customer the waitress had served.

b. The restaurant owner forgot which waitress the customer had served.

(34) a. The superintendent overheard which tenant the landlord had evicted.

b. The superintendent overheard which realtor the landlord had evicted.

Vincentile plots for the read-aloud task in Figure 48 show that RTs for the substitution/high-cloze condition (e.g., (34a)) were consistently faster than RTs for the substitution/low-cloze condition (e.g., (34a)) across bins. The statistical test corroborated this pattern. These show that verbs in the substitution/high-cloze condition were pre-activated more than verbs in the substitution/low-cloze condition. On the other hand, the same vincentile plots in Figure 48b do not show a consistent contrast across bins between the reversal/high-cloze condition (e.g., 33a) and the reversal/low-cloze condition (e.g., 33b), but only in the later bins. This was supported by the statistical analysis. These suggest that role-inappropriate verbs were pre-activated as strongly as role-appropriate verbs, which is consistent with the symmetric

pre-activation hypothesis.

There were increasingly slower RTs in the later bins in the reversal/low-cloze condition compared to the reversal/high-cloze condition. Although this was not the primary focus of this study, this pattern suggests that post-lexical processes slowed down RTs for the reversal/low-cloze condition.

Thus the two experiments showed different patterns. The high-cloze vs low-cloze contrast shifted the RT distributions of the reversal items in the lexical decision task, but not in the read-aloud task. The lexical decision task suggests that role-appropriate verbs are more strongly pre-activated than role-inappropriate verbs, but the read-aloud task suggests that their pre-activations are matched.

Surprisingly slow real-word responses in the lexical decision task suggest that the results from this task might be misleading, which might explain the mismatch between the tasks. As Figure 44a shows, the RTs for experimental items, where all targets were real words, were slower than the RTs for the OSV fillers, where all targets were non-words (pseudo-words). A post-hoc analysis with a linear mixed-effects model of lexical decision RTs contrasted real-word responses against non-word responses, and showed that real-word responses were significantly slower than non-word responses ($\beta = 228.94, p < .001$). That is, responses for real words were on average about 230ms slower than responses for non-words. Such slow responses for real words in a lexical decision task were unexpected given prior studies, where responses for real words were about as fast as or faster than responses for pseudo-words (e.g., Ratcliff et al., 2004). Such a large delay for the real-word responses might reflect post-lexical processes associated with the combinatorial representations of the utterance meaning. The meanings of the real-word targets can be integrated into the utterance meaning representa-

tions, but it is impossible for the non-word targets. Therefore, the slow real-word responses in the lexical decision task might have been influenced by post-lexical integrative processes. If this is the case, the contrasts in the locations of the RT distributions might reflect exactly the post-lexical processes that I aimed to eliminate in this study.

5.4 Experiment 5-2

In order to address the concerns about the first lexical decision experiment, I conducted a replication of the lexical decision task that differed from the original study in two ways. First, there was a deadline for the lexical decision responses. Since post-lexical combinatorial processes are not required to perform the lexical decision task, imposing a strict deadline for the responses should decrease the influence of the optional processes. The deadline was determined based on the latencies of non-word responses in Experiment 5-1. Since people could not be integrating non-words into the utterance meaning representations, the response latencies for those responses should not involve post-lexical combinatorial processes. Since about 75% of the non-word responses in the OSV fillers had RTs shorter than 1000ms, I assigned a 1000ms deadline in this experiment.

Second, in the comprehension questions, participants rated plausibility by clicking on a 1-7 Likert scale instead of providing binary judgments by pressing keys. This was because some pilot participants reported that the lexical decision responses were interfered with by the subsequent comprehension questions in the first experiment. This could have been because both tasks involved binary decisions and participants used identical keys for the tasks. In fact, responses in the read-aloud task were much faster compared to the lexical decision task (Figures 40), presumably because the manner of the tasks was quite different from the

plausibility judgment task and hence there was minimal interference from the plausibility judgment task. Therefore, in this experiment, participants rated the plausibility by clicking on a scale instead of pressing either of two keys in order to confirm that they read the sentences while minimizing the risk of interference from the comprehension questions. Additionally, I decreased the number of trials with comprehension questions, as reported in the Methods section.

5.4.1 Methods

Participants

133 native American English speakers were recruited online in the same way as Experiment 5-1. Participants received monetary compensation. All participants who failed to meet the deadline in more than 33% of the trials, or who provided inaccurate responses in 33% of trials were rejected.

It was necessary to exclude a large number of participants due to poor performance and due to the possibility that there were many submissions from very few participants. 61 out of 133 participants reported that they are exactly 25 years old. Among those 25-year-old participants, 49 failed to meet the performance-based criteria, which accounted for 80.3% of all the 25-year-old participants. On the other hand, only three participants out of the remaining 72 participants (4.2%) failed to meet the criteria. While those 25-year-old participants had different IP addresses and IDs in Amazon Mechanical Turk, the large skew in the age distributions and the performance suggest that most of the submissions by the 25-year-old participants were from a single person or very few people who did not perform the task appropriately. Since there were still more 25-year-old participants compared to other ages

even after excluding the submissions with poor performance, I excluded all 61 25-year-old participants from this study. Additionally, five participants were excluded for poor performance in the comprehension questions in the filler items in the way described below. This left 64 participants (31 females and 33 males; average age = 41.5) for analyses.

Materials

The stimuli were identical to the lexical decision task in Experiment 5-1.

Procedure

The task was conducted in basically the same way as the lexical decision task in Experiment 5-1. However, there was a one-second deadline to make the lexical decision after the target word or non-word was presented. The target was presented on the screen for one second and the deadline was indicated by the disappearance of the target. Participants were instructed to produce responses as quickly as possible within the deadline. A blank screen was presented for 500ms, following the deadline. 50% of the trials with real word targets were followed by a comprehension question, where participants provided a response by clicking on a 1-7 Likert scale. This accounted for 25% of all the trials. The proportion was intentionally reduced compared to the previous lexical decision experiment where there were comprehension questions following 37.5% of the trials. The proportion was reduced in order to reduce the risk of interference from the comprehension questions.

Exclusion

I only included participants whose average rating for the plausible filler items was higher by

at least 1 point than the average rating for the implausible filler items. Five participants were excluded due to this criterion. Among the trials where the remaining 64 participants provided responses, trials with reaction times longer than 1200ms were excluded from all analyses. Note that responses produced within 200ms after the one-second deadline were used for the RT distributional analyses to preserve the tails of the distributions. 0.9% of lexical decision responses in the reversal and substitution items were removed by this process, leaving 3747 trials for analyses.

Data analysis

The RT data were analyzed in the same as the lexical decision task of Experiment 5-1. There were on average 15.1 observations in the lexical decision task for each experimental item for each condition.

5.4.2 Results

Comprehension questions

The ratings for comprehension questions are summarized in Table 2. While there seems to be a bias towards plausible judgments in either sentence type (reversal or substitution), participants showed a clear preference for the plausible conditions over the implausible conditions.

Cloze	High-cloze	Low-cloze
Reversal	6.49	3.53
Substitution	6.53	4.01
Fillers	6.65	2.72

Table 14: The plausibility rating of the experimental items in Experiment 5-2. Participants rated the sentences on a 1-7 Likert scale.

Lexical decision task

The accuracy of the lexical decision tasks in the experimental items is summarized in Table 13. The overall accuracy across experimental and filler items was 97.6%.

Cloze	High-cloze	Low-cloze
Reversal	97.4%	97.0%
Substitution	97.3%	96.6%

Table 15: The accuracy of the lexical decision task in Experiment 5-2.

Non-distributional analysis

Inaccurate responses and responses produced after the one-second deadline were removed from the non-distributional analyses of RTs. This process left 3559 trials. The RTs for the accurate lexical responses are shown in Figure 49. A linear mixed-effects model of RTs revealed a significant effect of cloze ($\beta = 18.56, p < .001$) but no significant effect of sentence type ($\beta = -5.28, p = .40$) and no interaction between the two effects ($\beta = -5.93, p = .37$).

Additionally, I contrasted the RTs for experimental items, where all the targets were real words, and the OSV fillers, where all the targets were non-words. The aim was to confirm that lexical decision latencies for readl-word targets were not delayed as significantly as was seen in the first experiment. A linear mixed-effects model showed that real-word responses were significantly slower than non-word responses ($\beta = 28.43, p < .001$). However, the coefficient was small compared to the same analysis conducted for the lexical decision task in Experiment 5-1 ($\beta = 228.94$).

Distributional analysis

For the RT distributional analyses I included the responses that were produced after the deadline but were produced within 200ms following the deadline for the RT distributional analyses. Excluding inaccurate responses, there were 3639 lexical decision responses for

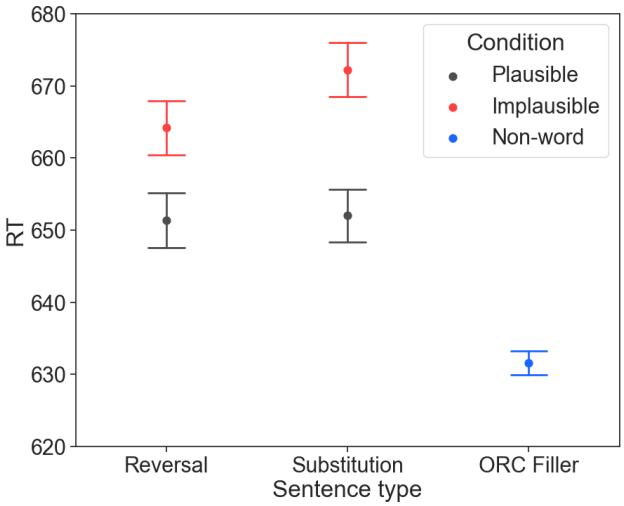
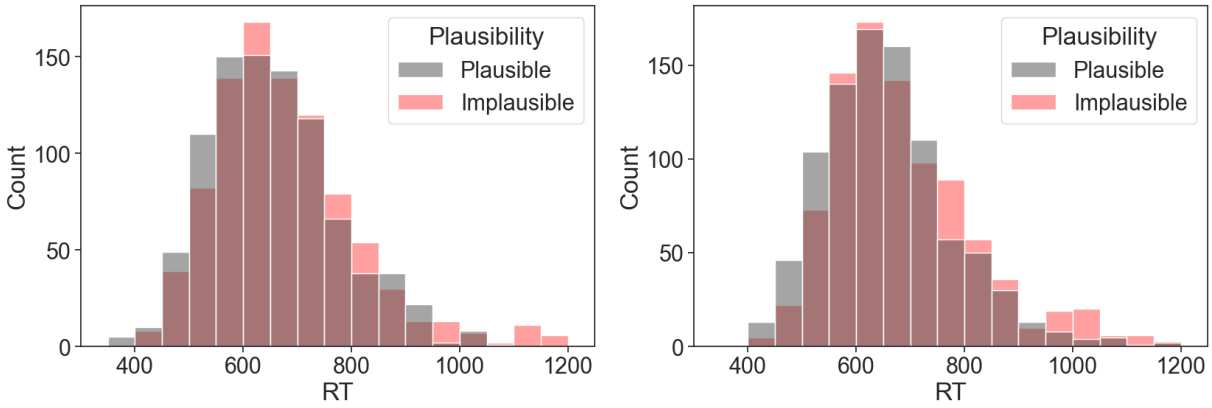


Figure 49: RTs of the experimental items (reversal and substitution) and the OSV fillers in the lexical decision task of Experiment 5-2. Error bars represent by-participants standard errors.

analyses. The RT distribution for each condition and each stimuli type is plotted in Figure 50.

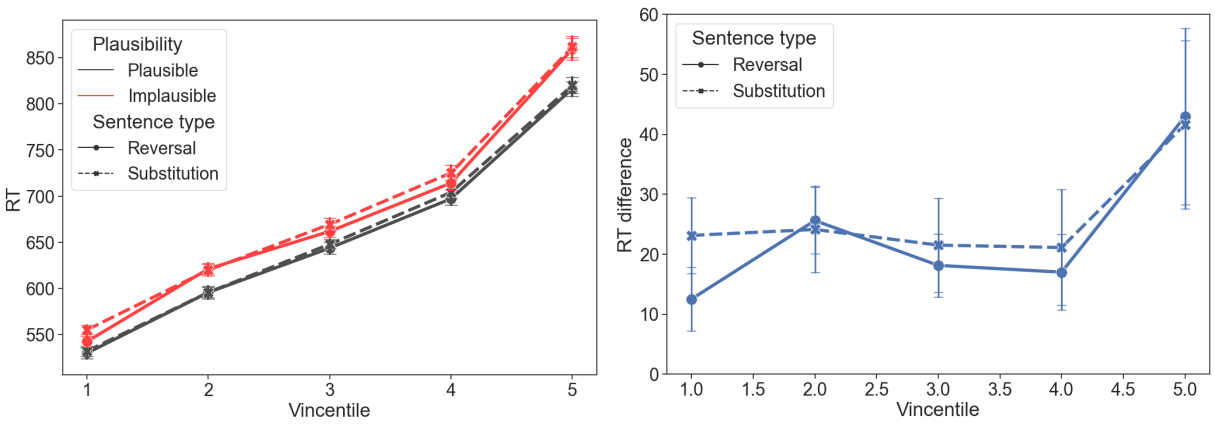


(a) Reversal **(b)** Substitution

Figure 50: RT distributions for the experimental items with accurate responses in the lexical decision task in Experiment 5-2.

The vincentiles are plotted in Figure 51a and the difference between the high-cloze and low-cloze conditions are plotted in Figure 51b.

A linear mixed-effects model of RTs revealed a significant effect of cloze ($\beta = 16.25, p < .001$), bin ($\beta = 262.95, p < .001$), and an interaction between the two factors ($\beta = 17.74, p =$



(a) Vincentile plots **(b)** Difference plots

Figure 51: Vincentile plots of the lexical decision task in Experiment 5-2. (b) shows the differences between the vincentiles of the original distribution (blue dots/lines in (a)) and the vincentiles of the modified distributions (orange and green dots/lines in (a)). Error bars represent the by-context standard errors.

.042). Other effects were not significant. This suggests that the locations of the RT distributions in the high-cloze condition are different from those of the low-cloze condition collapsing across the stimuli types (reversal vs substitution). The interaction between cloze and bin shows that this difference became greater in the slower bins. This suggests that the tails were also affected by the cloze manipulation. There was no evidence suggesting that the location or tail contrasts between the high-cloze and low-cloze conditions were different between the reversal and substitution items.

5.4.3 Discussion

This study aimed to conduct a lexical decision task similar to Experiment 5-1 but with reduced influence from post-lexical processes. In order to achieve this goal, participants were asked to provide responses within a one-second deadline. Additionally, the plausibility judgment in the comprehension questions in this experiment was collected by clicking on a Likert scale so that there was minimal risk of interference on the lexical decision task from

the comprehension questions.

The manipulations appear to have successfully reduced the involvement of post-lexical processes. One concern about the first lexical decision study is that the lexical decision responses for words in the experimental items were much slower than the responses for non-words in the OSV fillers, by about 200ms in Experiment 5-1. Such slower responses for words in the lexical decision task were unexpected since typically responses for words are faster than or about as fast as pseudo-word responses (e.g., Ratcliff et al., 2004). In Experiment 5-2, the RTs for the same experimental items and the same filler items were only different by 28ms, although the non-word responses were still faster.

The distributional analyses of RTs showed similar patterns as the lexical decision task in Experiment 1. As illustrated in the vincentile plots (Figure 51), the locations of the RT distributions of the high-cloze and low-cloze conditions were different both for the substitution items and the reversal items. There was no evidence for a difference between the substitution and reversal items. This suggests that role-appropriate verbs are pre-activated more strongly compared to role-inappropriate verbs.

5.5 General discussion

This chapter aimed to examine the pre-activation for role-appropriate and inappropriate candidates in argument role reversals. Findings from EEG studies using the N400 component (EEG study in Chapter 2; Kolk et al., 2003; Kuperberg et al., 2003; Hoeks et al., 2004; Kim & Osterhout, 2005; Oishi & Tsutomu, 2010; Chow, Smith, et al., 2016; Chow et al., 2018; Liao et al., 2022) as well as some eye-tracking studies (Kukona et al., 2011; Burnsky, 2022) have been taken to suggest that role-inappropriate candidates are pre-activated to a

similar extent to role-appropriate candidates, at least temporarily (symmetric pre-activation hypothesis). On the other hand, the speeded cloze study in Chapter 3 suggested that there is early sensitivity to argument roles, leading to large contrasts between the pre-activations of role-appropriate and inappropriate candidates (asymmetric pre-activation hypothesis). The mismatch between the pre-activation measures could be explained if measures showing sensitivity to argument roles in fact reflect post-lexical combinatorial processes. For example, as I proposed in Chapter 3, close measures might show sensitivity to argument roles because of a monitoring mechanism. Even when role-inappropriate candidates are pre-activated as strongly as role-appropriate candidates, the monitoring mechanism might detect the implausibility caused by the role-inappropriate candidates and inhibit them. Alternatively, the mismatch might also be explained if measures showing insensitivity to argument roles in fact reflect post-lexical combinatorial processes. For example, N400 might reflect the detection of semantic anomalies in the utterance representations as some researchers have argued (e.g., Kolk et al., 2003).

The current studies aimed to address the symmetric pre-activation hypothesis and the asymmetric pre-activation hypothesis by eliminating the influence of post-lexical processes, in order to obtain more direct evidence for either hypothesis. I conducted a lexical decision task and a read-aloud task. Neither of these tasks requires post-lexical processes to perform the task. Furthermore, I conducted distributional analyses of RTs in the tasks, in a further effort to dissociate the effect of pre-activation and post-lexical combinatorial processes. Differences in pre-activation should shift the locations of the RT distributions, as the lexical access processes should primarily affect basically all the trials. Such shifts in RT distributions can be observed as early and constant differences across bins (orange plots in Figure 42b). On

the other hand, differences in post-lexical processes should influence only the tails, because those processes should more strongly influence a subset of responses. This can be seen as no contrast in the fast bins but increasingly large contrasts in the slower bins (green plots in Figure 42b).

The lexical decision task and the read-aloud task in Experiments 5-1 and 5-2 showed clear parallelism in the substitution items. In both tasks, the substitution/high-cloze condition showed consistently short RTs from fast bins compared to the substitution/low-cloze condition. Hence the locations of RT distributions were not matched for the high and low-cloze conditions. This shows that the target verbs in the high-cloze condition were pre-activated more strongly compared to the target verbs in the low-cloze condition for the substitution items.

On the other hand, the two tasks did align with each other in the reversal items. In the read-aloud task, role-appropriate verbs in the reversal/high-cloze condition and role-inappropriate verbs in the reversal/low-cloze condition had matched RTs in the fast bins but increasingly large contrasts in the slower bins. Hence, the locations of their RT distributions were matched while the tails were not. The matched locations of RT distributions suggest that role-appropriate and inappropriate candidates have matched pre-activations. The results from the read-aloud task are consistent with Chow, Smith, et al., 2016's Experiment 1, which the stimuli of this study were taken from. Chow, Smith, et al. found N400 contrasts in the substitution items but not in the reverse items.

The lexical decision tasks in Experiments 5-1 and 5-2 consistently showed that role-appropriate verbs are produced faster than role-inappropriate verbs constantly from the fastest bins. Hence, the locations of the RT distributions were not matched for role-appropriate

and inappropriate verbs. This suggests that role-appropriate verbs are more strongly pre-activated compared to role-inappropriate verbs. The findings from the lexical decision tasks are more consistent with the speeded cloze task, where cloze responses show early sensitivity to argument roles.

What could have caused the discrepancy between the lexical decision task and the read-aloud task? In accounting for the mismatch between the tasks, it is important to highlight that the lexical decision task and the read-aloud task both agree that the high-cloze/substitution items and the low-cloze/substitution items do not have matched locations for their RTs. Hence both tasks suggest that pre-activations of the target verbs are not matched for these conditions. Therefore the cause of the discrepancy between the tasks must address a mismatch that is specific to the reversal items but not for the substitution items.

One simple explanation for the mismatch is that each task failed to show the effects that were found in the other task due to low statistical power. In fact, the visual pattern in the lexical decision tasks in Experiments 5-1 and 5-2 (Figures 46 and 51) suggest that the contrasts between the high-cloze and low-cloze condition of the reversal items might be smaller compared to the contrasts in the substitution items, although there was no support from the statistical analyses. However, it is somewhat unlikely that the lexical decision tasks in the two experiments consistently failed to find the contrast between the reversal and substitution items, although it is still possible.

Alternatively, the two tasks might have possibly eliminated the influence of post-lexical processes to different extents. Some researchers argue that lexical decision tasks are inherently more susceptible to post-lexical processes compared to read-aloud tasks (Seidenberg et al., 1984). If this is the case, post-lexical processes might influence most responses in the

lexical decision task and might shift the locations of RT distributions.

This chapter aimed to directly address the symmetric pre-activation hypothesis and the asymmetric pre-activation hypothesis, using two tasks and RT distributional analyses to isolate the effects of pre-activation from the effects of post-lexical processes. The different tasks yielded mismatching findings. The read-aloud task showed patterns consistent with EEG studies and suggests that role-appropriate and inappropriate candidates are activated similarly. On the other hand, the lexical decision tasks showed different patterns and suggest that role-appropriate and inappropriate candidates are pre-activated differently. Although the apparently conflicting findings require further investigation, the current studies suggest that the contribution of pre-activation and post-lexical combinatorial processes can be disentangled by distributional analyses of RTs.

6 The influence of role-inappropriate competitors on pre-activation

In Chapter 4, I explored the impact of role-inappropriate competitors (lures) on role-appropriate candidates through speeded cloze tasks. I demonstrated that contexts with stronger lures have longer role-appropriate RTs and I argued that this was because the strong role-inappropriate candidates inhibited the role-appropriate candidates. In this chapter, I turn to the potential impacts of role-appropriate competitors on role-inappropriate candidates.

Interestingly, there is a case where the presence of strong role-inappropriate competitors might influence role-sensitivity of pre-activation processes. Ehrenhofer et al. (2019) found N400 contrasts between role-appropriate and inappropriate verbs using stimuli such as (35) and (36). This is inconsistent with many other studies on argument role reversals using EEG and the N400 component (Kolk et al., 2003; Kuperberg et al., 2003; Hoeks et al., 2004; Kim & Osterhout, 2005; Oishi & Tsutomu, 2010; Chow, Smith, et al., 2016; Chow et al., 2018; Liao et al., 2022) or eye-tracking measures (Kukona et al., 2011; Burnsky, 2022). Ehrenhofer et al. also conducted an experiment where participans were presented with both Ehrenhofer et al.'s stimuli and the stimuli from a prior EEG study (Chow, Smith, et al., 2016), and Ehrenhofer et al.'s stimuli still elicited N400 effects whereas Chow, Smith, et al.'s stimuli did not.

Thus the N400 effects in Ehrenhofer et al.'s experiment could be attributed to the properties of the stimuli. Ehrenhofer et al.'s stimulus set was in fact unique in that contexts with both word orders were paired with highly predictable role-appropriate verbs (*ridden* in (35a) and *gored* in (36a)). Thus the role-inappropriate verbs in the argument role reversal

sentences ((35b) and (36b)) might be competing with strong role-appropriate competitors. Although the exact factor that contributed to the N400 effects in Ehrenhofer et al. (2019) is still unknown, the presence of such a strong role-appropriate competitor might have resulted in reduced pre-activation of the role-inappropriate verb, resulting in the N400 effects.

(35) a. The cattle rancher remembered which bull the cowboy had ridden out on the range.

b. The cattle rancher remembered which cowboy the bull had ridden out on the range.

(36) a. The cattle rancher remembered which cowboy the bull had gored out on the range.

b. The cattle rancher remembered which bull the cowboy had gored out on the range.

In this Chapter, I analyze speeded cloze data of both Chow, Smith, et al. (2016) and Ehrenhofer et al. (2019)'s stimuli collected in Chapter 4. I demonstrate that speeded cloze measures in fact provide evidence for stronger role-appropriate competitors in Ehrenhofer et al. than in Chow, Smith, et al.. I argue that such effects of stronger role-appropriate competitors are inconsistent with the asymmetric pre-activation hypothesis. On the other hand, I propose that the N400 effects in Ehrenhofer et al. (2019) can be explained by the sigmoid-shaped lateral inhibition and partial role-sensitivity in pre-activation mechanisms under the symmetric pre-activation hypothesis.

6.1 Ehrenhofer et al. (2019)

Ehrenhofer, Lau, and Phillips (2019) (henceforth ELP) used stimuli such as (35) and (36) in their EEG studies. The target verbs were high-cloze continuations for contexts with one word order and low-cloze continuations for contexts with the other word order. For example, in (35), *ridden* is a role-appropriate verb for (35a) but it is inappropriate for (35b). Importantly, their study was different from other prior studies in that there were two target verbs for each pair, and the verbs were appropriate in different sentences (e.g., *ridden* for (35a) and *gored* for (36a)). Therefore, there were two contexts and each of them was paired with a role-appropriate verb and a role-inappropriate verb. The target verbs were matched in offline cloze probabilities both as role-appropriate continuations and as role-inappropriate continuations. These stimuli were different from stimuli used in other prior studies such as (37), where researchers were not concerned with what role-appropriate verbs for (37b) would be.

(37) a. The old widower remembered which villager the ghost had haunted for many years.

b. The old widower remembered which ghost the villager had haunted for many years.

In Experiment 1, ELP presented these stimuli and unexpectedly found N400 contrasts between the sentences with role-appropriate verbs ((35a) and (36a)) and those with role-inappropriate verbs ((35b) and (36b)). These findings stand in contrast with the many other EEG studies that did not find contrasts between N400 amplitudes for role-appropriate and inappropriate verbs (e.g., (37a) and (37b)). Subsequently, in Experiment 2, ELP replicated

their findings from Experiment 1 as well as those from other prior studies. ELP took stimuli from Chow, Smith, Lau, and Phillips (2016)'s Experiment 1 (CSLP) such as (37), where no N400 effects were found for argument role reversals, and presented those stimuli together with their own stimuli such as (35) to the same participants in the same experiment. Henceforth, I will refer to the contexts where the target verbs were appropriate as the "canonical" items ((35a), (37a)) and those where the target verbs were inappropriate as the "reversed" items ((35b), (37b)). Ehrenhofer et al. found that there were N400 contrasts between the canonical and reversed item in ELP's stimuli but no such contrasts in CSLP's stimuli,

In order to identify the sources of the contrasts between the stimuli sets, ELP compared their stimuli with CSLP's stimuli in post-hoc analyses. In the offline cloze experiments, ELP did not find large contrasts in the cloze probabilities of the target verbs between the two sets of stimuli. The cloze probabilities were 28% in the canonical items and 1.1% in the reversed items for ELP's stimuli used in their Experiment 2, and 28% in the canonical items and 0.4% in the reversed items in CSLP's stimuli. There was no contrast in the entropies of the cloze responses either. The canonical and reversed contexts seemed to be equally constraining in CSLP and ELP, according to the offline cloze probability measures.

However, ELP's post-hoc analyses used offline cloze data, but speeded cloze data may reveal contrasts between the two stimulus sets that the offline cloze measures fail to capture. As shown in Chapter 2, cloze RTs can reveal some aspects of pre-activation that cloze probabilities cannot reveal. For example, while cloze entropies better reflect the dominance of a lexical candidate, mean RTs of cloze responses reveal the availability of lexical candidates. Even though ELP did not find contrasts in cloze entropy in CSLP and ELP, suggesting that there were equally dominant candidates, there could be contrasts in the availability of

role-appropriate candidates between the two stimulus sets, which is better reflected in cloze RTs.

6.2 Analyses of speeded cloze data contrasting CSLP and ELP

In order to examine the difference between CSLP and ELP's stimulus sets, I analyzed the speeded cloze data collected for these stimuli in Chapter 4. The details of the experiments are described in Sections 4.1 and 4.2. Those speeded cloze data were from two experiments with different sets of subjects and slightly different instructions. However, as I demonstrated in Chapter 4, participants responded very similarly to the common filler items shared between the two experiments. Therefore, the difference in the participants or instructions should have minimal effects on the results of the experiments.

Here I focus on ELP's Experiment 2, where ELP's stimuli were directly compared against CSLP's stimuli within a single experiment. In Experiment 2, ELP used 54 experimental items from each of the original sets of stimuli in CSLP and in Experiment 1 of ELP, instead of using the full 60 items from each set. Since it was not clear which items were removed from the original sets in ELP's Experiment 2, I analyzed all the stimuli from each experiment, excluding 4 experimental items that appeared in both experiments.

As mentioned above, ELP paired both canonical and reversed items with role-appropriate verbs that have high cloze probability. ELP also ensured that those verbs were role-inappropriate and had low cloze probabilities with the alternative word order. While ELP fully crossed the stimulus type (canonical or reversed) and the appropriateness of the verbs in Experiment 1 (e.g., using four sentences in (35) and (36)), they selected one pair of role-appropriate and inappropriate sentences with the same target verb from each set of four items (e.g., just using

(35) but not (36)).

I compared the target verb cloze probabilities, the modal cloze probabilities, the cloze entropies, and the mean RTs of all role-appropriate responses. I obtained those measures from the canonical and reversed items from each stimulus set. All the analyses reported below used linear mixed-effects models of one of the cloze measures. The models had context type (canonical and reversed; coded as -0.5 and 0.5), stimulus set (CSLP and ELP; coded as -0.5 and 0.5), and the interaction between the two factors as predictors. The models included the random intercept of context.

It should be noted that the cloze probabilities are calculated differently in my speeded cloze studies and the offline cloze studies in ELP. While I directly used the responses from participants, synonymous responses were collapsed and counted as the same responses in ELP.

6.2.1 Target probability

First, I analyzed the cloze probabilities of the high-cloze target items in each set of stimuli. For example, the cloze probabilities for *haunted* in (37) and for *ridden* in (35). The target cloze probabilities of each set of stimuli are plotted in Figure 52. On average, the target cloze probabilities for CSLP's canonical stimuli and reversed stimuli were 21.0% and 5.8%. On the other hand, they were 32.9% in the canonical word order and 4.0% in the reversed word order in ELP's stimuli. A linear fixed-effects model revealed significant effects of context type ($\beta = -0.22, p < .001$), stimulus set ($\beta = 0.069, p = .0030$), and the interaction between the two ($\beta = -0.10, p = .0016$). Target items were produced more frequently in the canonical items of ELP compared to the canonical items of CSLP. This could be revealing the choice of

the stimulus in ELP's Experiment 2, because some items with high-cloze target items were removed in this experiment in order to match the target cloze probabilities with the target cloze probabilities of CSLP.

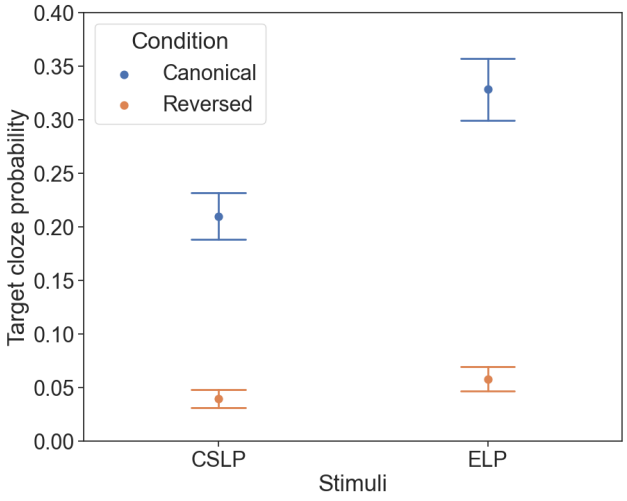


Figure 52: The target cloze probability in each stimulus set. The error bars indicate by-item standard errors.

6.2.2 Modal cloze probability

The modal cloze probabilities in each condition in each stimulus set are plotted in Figure 53. Visually, the canonical items seem to have greater modal cloze probabilities compared to the reversed stimuli, and the ELP items seem to have greater modal cloze probabilities compared to the CSLP items. A linear-mixed effect model of entropy revealed a significant effect of context type ($\beta = -0.065, p = .0030$) and stimulus set ($\beta = 0.048, p = .031$), but there was no significant interaction between condition and stimulus set ($\beta = -0.037, p = .39$). This confirmed the pattern seen in Figure 53.

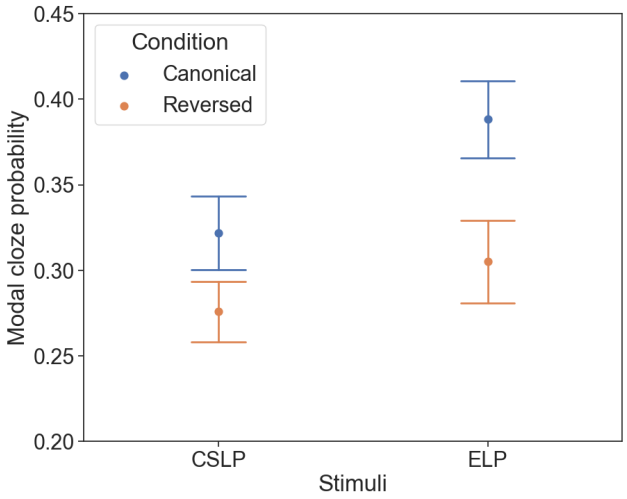


Figure 53: The modal cloze probability in the experimental items in each condition in each stimulus set. The error bars indicate by-item standard errors.

6.2.3 Entropy

The entropy of cloze responses was calculated for each item in the CSLP and ELP. The mean entropy for each condition for each stimulus set is plotted in Figure 54. Visually, the canonical items seem to have lower entropy compared to the reversed stimuli, and the ELP items seem to have lower entropy compared to the CSLP items. A linear-mixed effect model of entropy revealed a significant effect of context type ($\beta = 0.16, p = .0088$) and stimulus set ($\beta = -0.16, p = .029$). There was no significant interaction between condition and stimulus set ($\beta = 0.027, p = .029$). This confirmed the pattern seen in Figure 54. There was no evidence showing that the contrast between the canonical and reversed items was different between the stimulus sets.

6.2.4 Mean role-appropriate RTs

The mean role-appropriate RTs in each condition in each stimulus set are plotted in Figure 55. Visually, the canonical items seem to have shorter mean RTs for role-appropriate responses

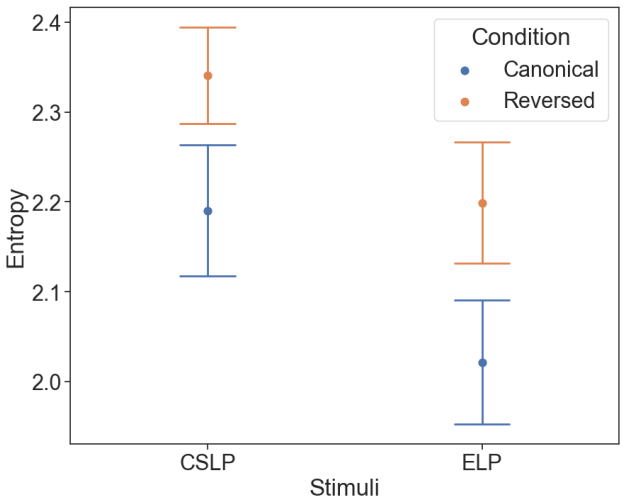


Figure 54: The entropy of the cloze responses in each condition in each stimulus set. The error bars indicate by-item standard errors.

compared to the reversed stimuli, and the ELP items seem to have shorter RTs compared to the CSLP items. A linear-mixed effect model of mean role-appropriate RTs revealed a significant effect of condition ($\beta = 106.71, p < .001$) and stimulus set ($\beta = -75.84, p = .0015$), but no significant interaction between condition and stimulus set ($\beta = 48.63, p = .30$). This confirmed the pattern seen in Figure 55.

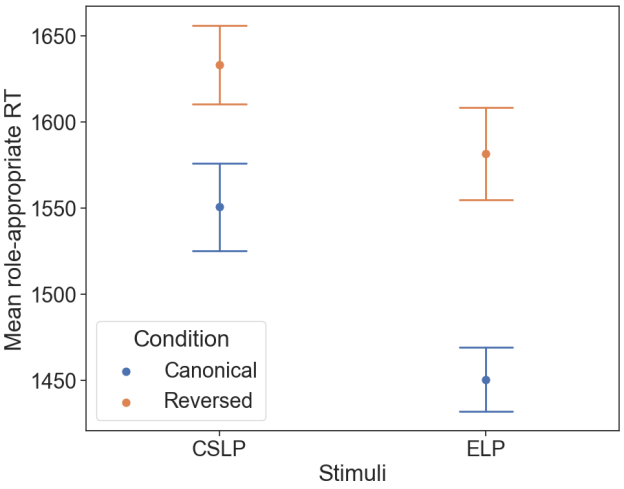


Figure 55: The mean role-appropriate RTs in the experimental items in each condition in each stimulus set. The error bars indicate by-item standard errors.

6.2.5 Discussion

ELP's analyses of offline cloze measures did not provide evidence for the key differences between the stimulus set of CSLP and ELP that can account for the presence/absence of N400 effects in the two experiments. However, the speeded cloze data suggest that the stimuli of CSLP and ELP were indeed different. Especially the contrasts in the mean role-appropriate RTs suggest that there were stronger role-appropriate candidates available in both the canonical and reversed items in ELP compared with CSLP.

Thus the stimuli were different from each other in at least these two aspects. First, ELP had stronger verbs appropriate in the canonical contexts. For example, *ride* in (39) was stronger than *haunt* in (38), and hence more strongly pre-activated. Second, ELP had stronger verbs appropriate in the reversed contexts. For example, *gore* in (39) were more strongly pre-activated than *witness* in (38).

(38) a. The old widower remembered which villager the ghost had haunted for many years.

b. The old widower remembered which ghost the villager had haunted for many years.

(39) a. The cattle rancher remembered which bull the cowboy had ridden out on the range.

b. The cattle rancher remembered which cowboy the bull had ridden out on the range.

The current study does not provide direct evidence for how these two factors contribute to the results, but prior evidence suggests that the strengths of verbs that are appropriate

for canonical contexts (*gore* vs *haunt*) are less important compared to the strengths of verbs that are appropriate in the reversed contexts (*ride* vs *witness*). Chow et al. (2018) showed in one of their EEG experiments in Mandarin Chinese that verbs that are highly likely in the canonical contexts (88% offline cloze probability on average) did not show N400 contrasts between the canonical and reversed contexts, just like CSLP. This study only used offline cloze data, but these high-cloze verbs should be strong candidates. Therefore, the presence of strong role-appropriate candidates in the canonical condition (e.g., *ride* for (39a) vs *haunt* for (38a)) itself is unlikely to explain the presence of N400 contrasts in ELP. Instead, the availability of the role-appropriate candidates in the reversed contexts (e.g., *gore* for (39b) vs *witness* for (38b)) seems to be more important for the contrast between CSLP and ELP.

Importantly, these role-appropriate candidates in the reversed contexts such as *gore* or *witness* were not presented in ELP's Experiment 2. Even though participants could have predicted them in the reversed condition, they were only presented with role-inappropriate candidates such as *ride* or *haunt*. Therefore, the presence of N400 effects in ELP seems to be driven by candidates that were not presented in the experiment.

6.3 Possible accounts of CSLP and ELP findings

The analyses of the speeded cloze data suggest that the presence of strong role-appropriate competitors in ELP resulted in N400 effects. In this section, I explore the implications of this finding on the pre-activation processes. I first argue that the presence of N400 effects in ELP is inconsistent with the asymmetric pre-activation for role-appropriate candidates and role-inappropriate candidates, since the unshown competitors should not affect N400 amplitudes under this hypothesis. Furthermore, I also argue that the symmetric pre-activation hypoth-

esis and lateral inhibition cannot straightforwardly explain the contrasts between CSLP and ELP as well. However, I propose that an additional assumption of slight role-sensitivity in pre-activation can explain the N400 effects found in ELP.

6.3.1 Difficulties of the asymmetric pre-activation hypothesis

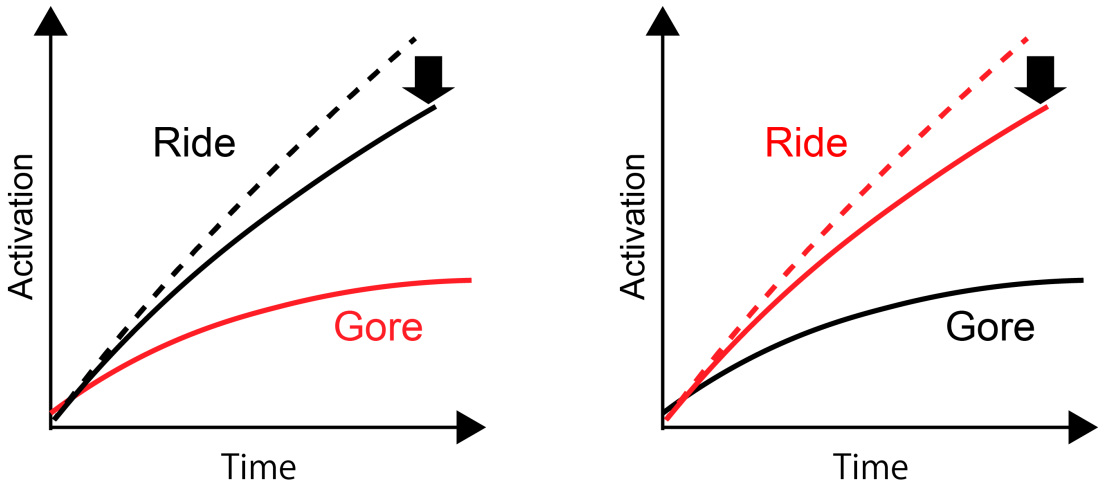
The presence of N400 effects in ELP is inconsistent with the asymmetric pre-activation hypothesis, because it questions the interpretation of the N400 component of the asymmetric pre-activation hypothesis. Under the asymmetric pre-activation hypothesis, pre-activation processes are role-sensitive and role-appropriate candidates are pre-activated more strongly compared to role-inappropriate candidates. As I discussed in Chapter 3, under this hypothesis N400 amplitudes show no sensitivity in many prior studies despite the contrast in pre-activation and hence it has to attribute N400 effects to post-lexical processes. For example, some researchers have argued that N400 amplitudes reflect the detection of semantic anomalies, and the absence of N400 effects in argument role reversals shows heuristic semantic processing that ignores the argument roles (e.g. Kim & Osterhout, 2005). For example, comprehenders might heuristically interpret (38b) as (38a).

However, it is difficult for such post-lexical interpretations of the N400 component to account for the influence of the unshown alternatives. As discussed above, the presence of role-reversal N400 effects in ELP seemed to be mostly driven by candidates such as *gore* that are not presented but would have been appropriate for the reversed contexts. If the absence of N400 effects in typical argument role reversal studies reflects semantic anomaly caused by role-inappropriate lexical items such as *ride*, unshown candidates (*gore*) should not be able to modify N400 effects.

6.3.2 Constraints on the symmetric pre-activation hypothesis

How can the symmetric pre-activation hypothesis account for the N400 effects found in ELP, presumably caused by unshown role-appropriate competitors? Lateral inhibition was one mechanism that I introduced in order to account for the effect of competitors. As I demonstrated in Chapter 4, lateral inhibition can explain how unspoken competitors influence the latency of responses in the speeded cloze task. The role-appropriate candidates competing with strong role-inappropriate candidates are inhibited by the role-inappropriate candidates, and as a result, the role-appropriate candidates are produced with longer RTs.

However, mere lateral inhibition cannot explain the N400 effects in ELP because ELP compares the same verb competing with the same competitors, whereas Chapter 4 compared different candidates competing with different competitors. If lexical candidates are pre-activated regardless of the roles as the symmetric pre-activation hypothesis posits, the same set of lexical candidates should be pre-activated to the same levels of activation in the canonical and reversed contexts. Therefore, both the target verbs and their competitors must have identical levels of pre-activation across the conditions in ELP. Then the target verbs should receive the same amount of inhibition in the canonical and reversed conditions. For example, a role-appropriate candidate *gore* for (39b) might inhibit role-inappropriate *ride* strongly, but at the same time role-inappropriate *gore* for (39a) must inhibit role-appropriate *ride* as strongly (Figure 56). The N400 amplitudes for the target verbs in the canonical condition must be identical to the N400 amplitudes for the target verbs in the reversed condition, reflecting the matched pre-activation levels and matched lateral inhibition from the competitors. Hence, mere lateral inhibition would not explain the contrasts in the N400 amplitudes



(a) Canonical condition **(b)** Reversed condition

Figure 56: Hypothetical pre-activation of the target verb *ride* in a stimulus in ELP (39) under the symmetric pre-activation hypothesis. Role-appropriate candidates are indicated by black lines and font color whereas role-inappropriate candidates are shown with red. The activation levels of *gore* are the same across conditions and hence the target verb *ride* receives the same amount of lateral inhibition. Dashed lines represent hypothetical activation patterns of the target verb without lateral inhibition.

between the target verbs in the canonical and reversed conditions.

In order to account for the lack of N400 effects in CSLP and the presence of N400 effects in ELP, I propose that there is partial role-sensitivity in pre-activation mechanisms. Role-appropriate candidates are slightly more pre-activated compared to role-inappropriate candidates. Combining this new proposal and the sigmoid lateral inhibition proposed earlier in Chapter 2, I argue that N400 effects in ELP arise from greater inhibition on the target verbs in the reversed condition compared to the canonical condition.

I propose that role-appropriate candidates are slightly more strongly pre-activated compared to role-inappropriate candidates. The difference in pre-activation levels is not large enough to elicit significant N400 effects, but the contrast between role-appropriate and inappropriate candidates can be exaggerated by lateral inhibition. In the reversed condition of ELP, role-appropriate *gore* in the reversed condition is strongly pre-activated compared

to role-inappropriate *gore* in the canonical condition. Therefore, the target verb *ride* is more strongly inhibited in the reversed condition than in the canonical condition. This contrasts in the pre-activation levels of *ride* results in N400 contrasts found in ELP.

Similarly, in CSLP, role-appropriate *witness* in the reversed condition is slightly more strongly pre-activated compared with role-inappropriate *witness* in the canonical condition. However, the higher levels of pre-activation for the competitors in the reversed condition do not result in significantly stronger inhibition on the target *haunt* due to the sigmoid inhibition function introduced in Chapter 2 (Figure 57). If the amount of lateral inhibition is determined by the sigmoid inhibition function, additional pre-activation for minimally-activated candidates does not result in greater inhibition (left side of Figure 57). Therefore, the contrasts in the pre-activation levels target verbs such as *haunt* is smaller than the target verbs in ELP and therefore do not yield N400 effects found in CSLP.

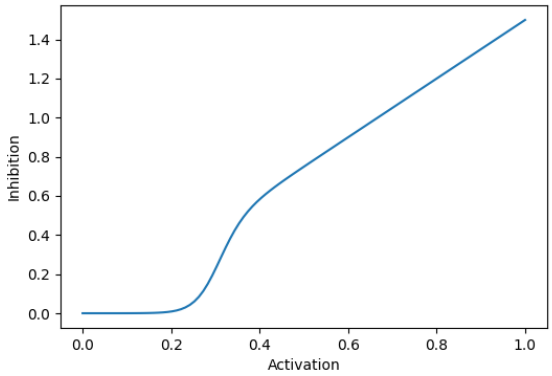


Figure 57: The function of the amount of inhibition each item imposes on its competitors introduced in Chapter 2.

This sigmoid inhibition was independently proposed and supported by Chen and Mirman (2012). They found a pattern across lexical access tasks and modalities that responses for target lexical items are slow when they have many phonologically or semantically associated

lexical items ("neighbors") that are strongly activated. On the other hand, responses for target lexical items with many weakly activated neighbors were faster. They argued that a sigmoid lateral inhibition function can successfully explain this pattern. In their account, neighbors mutually activate each other because their activation spread to each other via shared semantic or phonological representations. At the same time, they inhibit each other via lateral inhibition. With a sigmoid lateral inhibition function, weakly activated neighbors only very weakly inhibit the competitors. In this case, spreading activation from weakly activated neighbors to the target lexical items are greater than the inhibition from the neighbors. Therefore, weakly activated neighbors facilitate responses for the target items. On the other hand, the sigmoid lateral inhibition from strongly activated neighbors is stronger than the spreading activation, and hence strongly activated neighbors slow down responses for the target items.

There is another possible account in the literature proposed by Kuperberg (2016). Under this account, people decide whether or not to use argument roles to pre-activate lexical items depending on the availability of role-appropriate candidates. Comprehenders use argument roles for lexical prediction when there are strong role-appropriate candidates (e.g., *gore* in the reversed condition), and hence there are N400 effects found in ELP. On the other hand, comprehenders do not use argument roles when such strong role-appropriate candidates are absent in the reversed condition of CSLP, resulting in the lack of N400 effects.

In Kuperberg's probabilistic model, people generate expectations about the likely continuations based on multiple cues, such as the full combination of arguments and roles {villager-agent, ghost-patient} or just the pair of the arguments without roles {villager, ghost} [6] . In

[6]Note that this is a much-simplified explanation of Kuperberg's account. Kuperberg's model is a hierar-

CSLP's reversed condition, the full set of {villager-agent, ghost-patient} does not yield any dominant candidate, and hence the predictions are not constraining. However, the pair of {ghost, villager} might yield more constraining predictions where *haunt* is dominant. In such a situation, people deem the cues that yield more constraining predictions (the argument pair without roles) to be more "reliable" as a cue, and hence only rely on them. The reliability of cues is different in ELP stimuli. In the reversed condition of ELP, {bull-agent, cowboy-patient} already yields a dominant candidate *gore*. On the other hand, the role-less combination of {cowboy, bull} may yield less constraining predictions including both *gore* and *ride* as salient candidates. In such cases, people deem the combinations of arguments and roles to be more reliable. As a result, people use the combinations of arguments and their roles to generate predictions, and hence *gore* is strongly pre-activated but *ride* is not. Thus, Kuperberg's account maintains that the presence of a strong role-appropriate candidate blocks the role-insensitive predictions, resulting in N400 effects. This account successfully predicts the contrast in N400 effects in CSLP and ELP.

However, in order to identify the most reliable set of cues as Kuperberg argies, comprehenders must generate multiple possible pre-activation patterns generated by different sets of cues and compare them. Given a number of the types of potentially useful cues, it seems to be computationally highly costly to generate and compare multiple sets of pre-activation using each combination of the cues. Additionally, Kuperberg's account has difficulty in explaining the lack of N400 effects in the Japanese EEG experiment in Chapter 3. Since the contexts were as simple as just a single case-marked noun phrase (e.g., *bee*-NOM), it is unlikely that

chical generative model, and comprehenders generate expectations about high levels of representations, such as messages. Then the expectations about messages influence expectations about the lower levels, such as upcoming lexical items.

just a single noun phrase ignoring the case maker would result in more constraining predictions (also see Chapter 3).

In this Chapter, I examined the influence of role-appropriate competitors by contrasting two studies, CSLP and ELP. Despite the robust findings in EEG studies showing the lack of N400 effects in argument role reversals, which was corroborated by CSLP, ELP found N400 effects for argument role reversals. The current analyses of the speeded cloze data addressed what differences in the stimuli caused the N400 effects only in ELP, and showed that there are strong role-appropriate candidates available in the reversed condition in ELP but not in the same condition in CSLP. Although the asymmetric pre-activation hypothesis has difficulties in accounting for such effects of unshown role-appropriate candidates in EEG studies, the symmetric pre-activation hypothesis might be able to account for the N400 effects in ELP by small role-sensitivity and the sigmoid lateral inhibition function.

7 Synthesis of the findings

In this thesis, I have investigated the competitive dynamics of lexical prediction mechanisms. I used the Simple Activation Model as a working model, which consists of the pre-activation mechanisms that map contextual information and knowledge to lexical pre-activation, and the theories of measures that map lexical pre-activation onto observable experimental measures. The investigation of the linking hypotheses of cloze measures and the studies on the role-(in)sensitivity of pre-activation processes provided substantial insights into the predictive processes. In this chapter, I first present a summary of the findings from each study in this thesis, and subsequently synthesize the findings regarding how they update the Simple Activation Model.

7.1 Summary of the thesis

In Chapter 2, I examined the Race model (Staub, 2011), which is a linking hypothesis for the speeded cloze task. Under the model, cloze responses are the first lexical candidates whose pre-activations reach a threshold before their competitors (i.e. winners of "races"). I demonstrated that although the original model can capture the basic relation between cloze probabilities and response times, it struggles to capture the steep slope of this relation, i.e., rare responses are produced more slowly than the model predicts. I demonstrated that lateral inhibition between lexical candidates, while maintaining other assumptions about the race, enables such slow wins. Subsequently, by analyzing simulated cloze data, I examined which cloze measures reflect which aspects of the pre-activation mechanisms. The analyses showed that cloze RTs mainly reflect the strength of lexical candidates (i.e., average rates of

pre-activation) while cloze probabilities better reflect the levels of pre-activation. Moreover, the mean cloze RTs for a context reveal the availability of candidates while the modal cloze probabilities or the entropies show how dominant one candidate is.

In Chapter 3, I examined argument role reversals, where different measures of pre-activation do not align with each other. Prior studies have found that N400 amplitudes seem to be insensitive to argument roles while offline cloze measures show a high degree of sensitivity to argument roles. I demonstrated that prior findings about the mismatch between cloze probabilities and N400 amplitudes cannot be attributed to timing contrasts. I re-analyzed an EEG study by Momma (2016) that used maximally simple Japanese stimuli such as *Bee*-NOM *sting* and I conducted a speeded cloze study that was matched with the EEG study in terms of the time courses and the stimuli. The studies showed that (i) the measures do not align with each other even when they probe pre-activation at the same moment, but (ii) N400 amplitudes do show sensitivity to argument roles when there is additional time for pre-activation. These two contrasts raised questions about whether and how pre-activation impacts argument roles, and about what different measures of pre-activation actually reflect.

In Chapter 4, I examined the evidence for some degree of role-insensitive pre-activation in speeded cloze tasks. I estimated the strengths of role-inappropriate candidates (lures) in each context using the speeded cloze RTs, and I showed that the context-wise variability of the estimated strengths of lures covaries with (i) the RTs of role-appropriate responses and (ii) the proportion of role-inappropriate responses. These are consistent with accounts where role-appropriate candidates are activated to the extent that they can inhibit role-appropriate competitors and win races at least occasionally.

In order to further examine the evidence for role-(in)sensitivity in pre-activation, Chapter 5 aimed to dissociate the influence of argument roles on lexical pre-activation from their effects on post-lexical processes. I used tasks that do not, in principle, require post-lexical combinatorial processes, and I used RT distributional analyses to examine role-sensitivity in lexical pre-activation excluding the effect of later processes. The read-aloud task suggested that role-appropriate and inappropriate candidates are pre-activated to a similar extent, whereas the lexical decision task suggested that role-appropriate candidates are pre-activated more strongly compared to role-inappropriate candidates.

Chapter 6 assessed the contribution of role-appropriate competitors on role-inappropriate pre-activation. I focused on findings that stimuli created by Ehrenhofer et al. (2019) can cause N400 effects in argument role reversals but those created by Chow, Smith, et al. (2016) do not. I analyzed speeded cloze data collected for the two sets of stimuli in Chapter 4, and demonstrated that the key difference between the two EEG studies was the availability of strong role-appropriate competitors. The asymmetric pre-activation hypothesis has difficulties in accounting for such effects of unshown role-appropriate candidates in EEG studies. On the other hand, the symmetric pre-activation hypothesis might be able to account for the N400 effects in ELP by the sigmoid lateral inhibition function and additional pre-activation for role-appropriate candidates.

7.2 Implications for the Simple Activation Model

Here I synthesize the findings in terms of how they allow us to update the Simple Activation Model (Figure 58). The Simple Activation Model is a model combining pre-activation mechanisms that map knowledge and contextual information onto lexical pre-activation, and

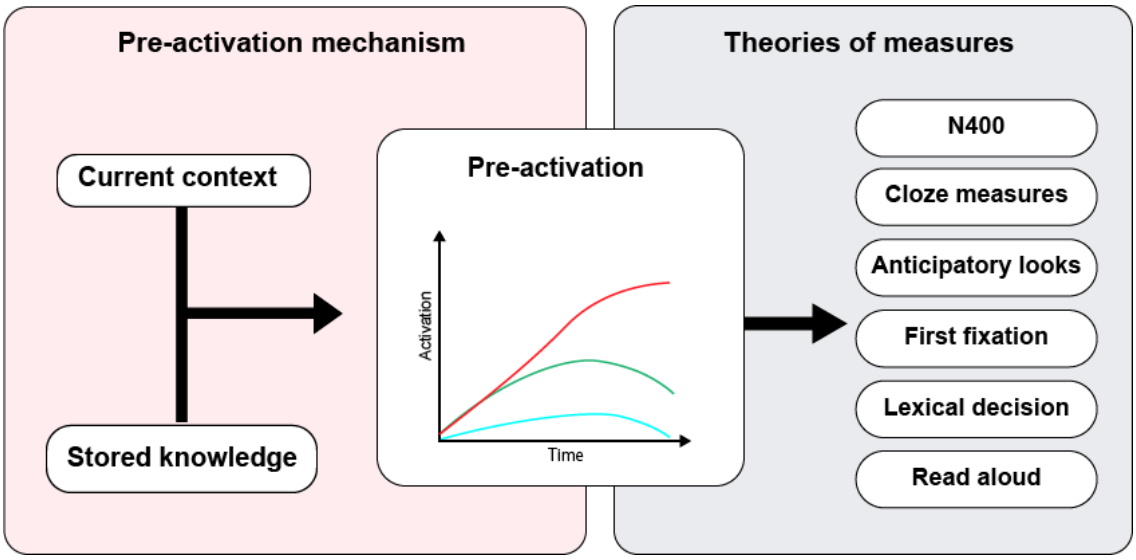


Figure 58: The Simple Activation model

a model of measures that maps lexical pre-activation values onto observable experimental measures. The key assumptions of the "simple" activation model are that there is a shared underlying pre-activation process and that it is transparently reflected in widely used experimental measures of pre-activation.

This thesis attempted to understand lexical prediction by updating this model. Argument role reversals are particularly useful test cases in this regard because they provide insights into both the pre-activation mechanism and the theories of measures. Pre-activation measures show sensitivity to argument roles in some cases but not in other cases. Knowing under what circumstances argument roles impact pre-activation would be informative about how the pre-activation mechanisms use contextual information and stored knowledge to generate pre-activations. The discrepancies between pre-activation measures are also informative about what cognitive processes the measures actually reflect. I summarize how the current thesis combined with prior studies allows us to update the pre-activation mechanisms and the theories of measures in the Simple Activation Model.

7.2.1 Shared pre-activation

7.2.2 Theories of pre-activation measures

The important and consistent issue for the measures of pre-activation is whether they reflect pre-activation, post-lexical combinatorial processes, or both. The current studies provide insights into which process the measures are sensitive to, and ways to dissociate the two factors when their effects are combined in the same measure.

Cloze measures

This study provided support for the Race model for the underlying processes of the speeded cloze task. While the original model by Staub et al. (2015) made quantitatively inaccurate predictions, the issue can be solved by assuming lateral inhibition in pre-activation, which has independent support from the bottom-up lexical activation literature.

The thesis also highlighted what aspects of pre-activation different cloze measures reflect. The analyses of simulated cloze data showed that cloze RTs are more sensitive to the rates of pre-activation while cloze probabilities better reflect the levels of pre-activation. Additionally, while previous studies have used the notion of "contextual constraint" to characterize the set of lexical candidates pre-activated by each context, the Race model suggests that this conflates the two different notions of the availability of candidates and the dominance of a candidate. These two notions can be separately measured via different cloze measures: the mean cloze RTs for availability and modal cloze probabilities or entropies for dominance.

The studies on argument role reversals suggested a new update in the cloze process. If role-appropriate and inappropriate candidates are similarly pre-activated, the frequent production

of role-appropriate candidates and the rare production of role-inappropriate candidates can be explained by a speech monitoring mechanism that prevents the production of implausible utterances (Levelt, 1989). Once a lexical candidate reaches a threshold, it is checked for its plausibility in the context. If the whole utterance becomes implausible because of the lexical candidate, it is inhibited to prevent implausible utterance. The same mechanism can also explain some N400 effects found in argument role reversals (Chapter 3).

N400 component of ERP

Some researchers have argued that the N400 component for a lexical item reflects predictive processes. Most of them assume that the N400 amplitudes reflect pre-activation levels of the lexical or lexico-semantic representations (e.g., Federmeier & Kutas, 1999; Lau et al., 2008), while some have argued that they instead reflect predictions in combinatorial utterance meaning representations (Rabovsky et al., 2018). Either account maintains that N400 is sensitive to processes that occur prior to the bottom-up processing of new inputs.

On the other hand, other researchers have argued that the process reflects more passive processes that follow the bottom-up processing of an incoming lexical item. Under these passive accounts of N400 amplitudes, some researchers argue that N400 amplitudes reflect the effort of building utterance meaning representations (e.g., Brown & Hagoort, 1993; Kolk et al., 2003; van Herten et al., 2005), while others interpret N400 amplitudes as signals of semantic anomalies in the constructed utterance meaning representations (e.g., Kim & Osterhout, 2005; Li & Ettinger, 2023).

The findings in the current thesis are more consistent with the predictive view than the passive view. Since the passive view maintains that the N400 absence in argument roles

reflects processes triggered by the presentations of the role-inappropriate continuations, it has difficulty in explaining cases where processes prior to the presentations or the lexical candidates that are not presented can "turn on" the N400 contrasts. As shown in Chapter 3, Momma (2016)'s EEG study showed sensitivity to argument roles when the target verb was presented with delays, consistent with prior studies (Chow et al., 2018; Liao et al., 2022). Since the passive accounts attribute the general N400 absence in argument role reversals to the processes following the presentation of the target lexical item, it is hard for such accounts to explain the influence of additional time prior to the presentation of the target item. On the other hand, a predictive account can more straightforwardly explain the timing effects by arguing that predictions change over time. One way for the passive accounts to explain these timing effects is to attribute them to dynamic changes in the context representations. The additional time prior to the target presentations might increase the time to process the contexts, potentially changing the context meaning representations. If this is the case, then having additional time to process the context might influence the construction of new utterance meaning representations based on the context meaning representations.

However, the N400 contrasts in Ehrenhofer et al. (2019) are more difficult for the passive accounts to explain. My analyses of the speeded cloze data for Ehrenhofer et al.'s stimuli suggested that the key factor that caused the N400 contrasts was the strength of lexical candidates that were not presented. If N400 amplitudes only reflect processes triggered by the presentation of the lexical items, other lexical items that are not presented should not influence the N400 amplitudes. On the other hand, if N400 amplitudes reflect predictive processes that occur prior to the presentation of the lexical item such as the amount of pre-activation, pre-activations for other items can influence the pre-activation for the presented

lexical item. Thus the predictive view of the N400 component is more consistent with the role-sensitivity of the N400 component caused by the additional time prior to the target presentation or by competitors that were predicted but not presented.

A related but different question is whether N400 reflects lexical (or lexico-semantic) processing or post-lexical combinatorial processes such as utterance meaning. This is related to the contrast between the predictive and passive distinction because the passive accounts attribute the N400 effects to combinatorial processes while most predictive accounts argue that it is at the lexical or lexico-semantic level. An exceptional account is Rabovsky et al. (2018), where N400 amplitudes reflect changes in predicted utterance meaning representations. The utterance meaning representations in Rabovsky et al.'s model are predictive in that the comprehender predicts the meaning of the entire utterance at any point as the utterance unfolds. If a new input is consistent with the current utterance meaning representation, it results in a small change in the utterance meaning representation, whereas inconsistent inputs require a large change in the utterance meaning representation. Rabovsky et al. argue that N400 amplitudes reflect such changes in the utterance meaning representations. However, Rabovsky et al.'s model assumes a transparent relationship between predictions for the utterance meanings and for upcoming lexical items. Therefore, although N400 amplitudes reflect post-lexical processes in Rabovsky et al.'s model, N400 amplitudes should strongly correlate with lexical pre-activation. Thus the current findings and prior literature support the view that N400 amplitudes either rather directly reflect lexical pre-activations or at least strongly correlate with them. Therefore I maintain that there is good justification for treating N400 amplitudes as pre-activation measures

Lexical decision task and read-aloud task

The lexical decision task and the read-aloud (naming) task can also be considered as pre-activation measures since the responses seem to be facilitated by lexical pre-activation via priming or sentential contexts (e.g., Stanovich & West, 1983; Seidenberg et al., 1984; Andrews & Heathcote, 2001; Ratcliff et al., 2004; Balota et al., 2008). They also correlate with (offline) cloze probabilities (Kleiman, 1980; Seidenberg et al., 1984).

Since these tasks do not require combinatorial processes following lexical access to the target items, responses to these tasks might be useful in assessing the pre-activation without the influence of the post-lexical processes. While it is still possible that post-lexical processes influence the responses in these tasks, RT distributional analyses can potentially discriminate the two processes. While pre-activation should affect all trials, post-lexical processes should only influence a subset of trials where they were deployed. Then the effect of pre-activation should appear mainly as a shift in the locations of RT distributions while post-lexical processes should only influence the tails of the distributions. However, the studies in Chapter 5 showed a puzzling mismatch between the two tasks, and therefore the linking hypotheses require more elaboration.

7.2.3 Theories of pre-activation generation

How candidates are pre-activated

The current thesis has implications for the way in which lexical candidates are pre-activated. The parallel pre-activation of predicted candidates was one of the assumptions of the Simple Activation Model, while some researchers pointed out the possibilities of more serial prediction in at least some tasks (Van Petten & Luka, 2012). The evidence for the Race model also

supports parallel activation. As Staub et al. (2015) argued, the lexical candidates competing with fast competitors have short RTs because they must be faster than those competitors in order to be produced. This is only possible when multiple candidates are pre-activated in parallel. This thesis provided additional support for a parallel pre-activation mechanism by reinforcing the support for the model.

This thesis also provided evidence for lateral inhibition between the simultaneously pre-activated candidates. As shown in Chapter 2, lateral inhibition can explain slow cloze responses that were unexpected by the original simulations by Staub et al. (2015). Furthermore, Chapter 4 showed that role-appropriate candidates competing with strong role-appropriate candidates have longer RTs. This shows that stronger role-inappropriate candidates inhibit the role-appropriate candidates strongly, resulting in slower role-appropriate RTs.

Impact of contextual information and knowledge: Argument roles

In addition to the general manners of pre-activation, this thesis also has implications on how contextual information and stored knowledge impact pre-activation. I paid special attention to argument roles and contrasted the pre-activation of role-appropriate and inappropriate candidates such as *serve* in (40). Even though argument roles are crucial in order to comprehend utterances, prior studies have shown that pre-activation measures are not sensitive to argument roles (N400: e.g., Kolk et al., 2003; Kim & Osterhout, 2005; Chow et al., 2018; Liao et al., 2022; anticipatory looks in the visual world paradigm: Kukona et al., 2011; eye-tracking while reading: Burnsky). However, this thesis has shown that argument roles can impact pre-activation measures at least under some circumstances. This raises a question of whether and how argument roles can impact pre-activation, and what is the nature of

pre-activation mechanisms that causes such patterns.

(40) a. The restaurant owner forgot which customer the waitress had served.

b. The restaurant owner forgot which waitress the customer had served.

The question of whether and how argument roles impact pre-activation can be broken down into smaller pieces. First, are argument roles available to comprehenders at all when they generate predictions? Second, can argument roles constrain pre-activation in some way? Finally, in what circumstances can argument roles constrain pre-activation? The findings of the current thesis can address each of the questions.

The first question is whether argument roles are available to any cognitive mechanism in argument role reversals at the moment pre-activation is generated. While comprehenders must identify argument roles at some point in sentence comprehension, the findings in the EEG studies that delayed presentation of the target verbs can turn on the N400 effects (Chapter 3) might be interpreted to show that argument roles are identified slowly. That is, when the target verbs are presented shortly after the contexts, the roles of the arguments in the contexts are not identified yet, and hence the continuations are processed similarly regardless of the argument role assignments. On the other hand, the roles can be identified given additional time, and therefore N400 amplitudes were sensitive to argument roles when the target verbs were presented following a longer delay.

However, the current studies provide evidence against such slow identifications of argument roles. In this thesis, there were some cases where people showed sensitivity to argument roles shortly after the context presentation (the Japanese speeded cloze study in Chapter 3 and the English lexical decision task in Chapter 5). Regardless of the exact mechanisms in

which argument roles impacted performance in these tasks, the role-sensitivity in these cases shows that argument role information is available to at least some cognitive mechanisms, even in the early moments of processing where N400 does not typically show role-sensitivity.

Given that argument roles are available in the comprehender's mind, a second question is whether argument roles can impact pre-activation in some way. The current thesis and prior studies suggest that argument roles do influence pre-activation. In addition to the speeded cloze studies and the lexical decision study that showed sensitivity to argument roles, there are some cases where N400 amplitudes show sensitivity to argument roles (Momma (2016) in Chapter 2 and Ehrenhofer et al. (2019) discussed in Chapter 6, as well as other related studies (Chow et al., 2018; Liao et al., 2022; Lee & Phillips, 2023)). While the individual linking hypotheses for these measures might be questioned, the role-sensitivity found across multiple measures that are considered to reflect pre-activation suggests that pre-activation can be sensitive to argument roles at least under some circumstances.

Finally, even though argument roles themselves are available in a relatively early moment and they can impact pre-activation in some cases, why do pre-activation measures show sensitivity to argument roles only under some circumstances? I offer three possibilities to account for what kind of pre-activation mechanisms can cause the role-(in)sensitivity. Since none of the three accounts successfully explains all the findings, I lay out what can and cannot be explained by each hypothesis.

Asymmetric pre-activation

A first possibility is the asymmetric pre-activation hypothesis, where argument roles influence pre-activation consistently and even in argument role reversals. Under this hypothesis, argu-

ment roles in the context are used successfully, just like other cues such as the identities of the arguments, and hence role-appropriate candidates have much greater pre-activation compared to role-inappropriate candidates. This hypothesis can easily explain the role-sensitivity found in certain pre-activation measures. For example, under this account, most responses in the speeded cloze tasks are role-appropriate because those role-appropriate candidates are pre-activated more than role-inappropriate candidates. Regarding the findings from Chapter 4 that showed evidence for some pre-activation for role-inappropriate candidates, there must be some mechanism that allows role-insensitive pre-activation, such as lexical association.

On the other hand, it is difficult for this hypothesis to explain the findings in EEG studies. There should be asymmetric pre-activation for role-appropriate and inappropriate candidates under this account, but the N400 amplitudes for those candidates are matched in most cases. Therefore this hypothesis needs to argue that N400 amplitudes do not reflect pre-activation, but instead reflect other processes such as post-lexical combinatorial processes. In fact, some researchers argue that the N400 component reflects the detection of a semantic anomaly in the utterance, and N400 amplitudes are not sensitive to argument roles because comprehenders construct heuristic semantic representations that are unfaithful to the inputs but are plausible (e.g., Kim & Osterhout, 2005; Li & Ettinger, 2023). For example, when presented with (40b), people interpret it as (40a), and hence there is no semantic anomaly.

However, as pointed out above, such an account of N400 amplitudes has difficulty in explaining the presence of N400 effects in argument role reversals when additional time is provided prior to the presentation of the target item (Chapter 3) or when there are strong role-appropriate competitors (Chapter 6).

A potential solution for this hypothesis is to take a predictive combinatorial account

such as Rabovsky et al. (2018), where N400 amplitudes reflect changes in the predicted utterance meaning representations. However, Rabovsky et al. (2018)'s model itself cannot explain the lack of N400 contrasts despite the contrast in pre-activation. Since the model still predicts that N400 amplitudes correlate with lexical pre-activation, if there is asymmetric pre-activation for role-appropriate and inappropriate candidates, there should be contrasts in N400 amplitudes as well. Therefore, the asymmetric pre-activation hypothesis needs a new linking hypothesis for the N400 component where N400 amplitudes reflect a combinatorial process that is predictive but is distinct from lexical pre-activation at least in argument role reversals.

Symmetric pre-activation: Format mismatch

Another two hypotheses are versions of the symmetric pre-activation hypothesis, according to which role-appropriate and inappropriate candidates are similarly pre-activated. One of the hypotheses attributes role-insensitivity to mismatch between the format of the representation of the contextual information and the format of long-term knowledge, which are the two inputs to the pre-activation mechanisms in the Simple Activation Model (Figure 58). Chow, Momma, et al. (2016) argue that lexical prediction using event knowledge is a memory retrieval process where events consistent with the input are retrieved from long-term memory. Chow, Momma, et al. further argue that pre-activation might be insensitive to argument roles because the format of the argument roles as cues for memory retrieval does not match the representation of the event knowledge. Chow, Momma, et al. propose that comprehenders may encode the arguments in an utterance with their generic roles such as *waitress*-agent and *customer*-patient with the absence of the verb, whereas event representations in long-

term memory are instead represented in event-specific formats such as *waitress*-server and *customer*-servee. Since the available cues from the generic argument roles do not match the target representations of event-specific roles, those argument roles are not effective cues for memory retrieval. Thus, comprehenders might have access to events that involve the arguments but not those that involve the arguments with specific roles. As Liao et al. (2022) argue, pre-activation via such a role-insensitive memory search might be preceded by pre-activation via simple association, which makes the format mismatch account of the absence of N400 effects consistent with other accounts attributing the absence of N400 effects to associative pre-activation (e.g., Brouwer et al., 2012).

The format mismatch account can thus explain the insensitivty to argument roles, but it also has to account for three cases discussed in this thesis where pre-activation measures show sensitivity to argument roles: (i) the speeded cloze task (Chapter 3), (ii) N400 amplitudes with delayed presentation of the target items (Chapter 3), and (iii) N400 amplitudes with strong role-appropriate competitors in Ehrenhofer et al. (2019) (Chapter 6). These exceptional cases can be explained by additional mechanisms.

I proposed a monitoring mechanism to account for (i) and (ii). In the speeded cloze task, once a lexical candidate's pre-activation reaches a threshold, a monitoring mechanism checks the plausibility of the entire utterance including the lexical candidate. When the utterance would be inappropriate, the lexical candidate is inhibited. Role-inappropriate candidates are detected and inhibited before they are produced by this mechanism. I further proposed that the monitoring mechanism is also involved in comprehension, and that additional time for pre-activation enables more successful detection of role-inappropriate candidates. This can explain the findings in the timing contrasts in the Japanese EEG experiment by (Momma,

2016) (Chapter 3) as well as other converging prior studies (Chow et al., 2018; Liao et al., 2022). The earlier sensitivity in the speeded cloze responses compared to N400 amplitudes can be explained by differences in the threshold. The threshold is lowered in the speeded cloze task because people need to ensure that at least one candidate reaches the threshold and can be produced quickly. Thus even when role-inappropriate candidates are highly pre-activated due to role-insensitive memory retrieval, they can be inhibited by the role-sensitive monitoring mechanism. This hypothesis is in fact consistent with Lee and Phillips (2023), where they observed N400 effects for argument role reversals in an EEG design where speeded cloze trials were interleaved with comprehension trials. The monitoring hypothesis predicts this finding because the interleaved cloze trials would lower the threshold for the monitoring process, which enables early detection of role-inappropriate candidates in the other trials.

The role-sensitivity of N400 amplitudes found in Ehrenhofer et al. (2019) can potentially be explained by small additional pre-activation for role-appropriate candidates. The additional role-sensitive pre-activation is small and it does not cause N400 effects for argument role reversals by itself. However, the small contrast can be exaggerated by lateral inhibition in Ehrenhofer et al.'s experiment. The additional pre-activation for role-appropriate competitors in Ehrenhofer et al. results in stronger inhibition for the role-inappropriate target items in the reversed condition. Therefore the target verbs have significantly different levels of pre-activation in the canonical and reversed conditions and elicit greater N400 amplitudes in the reversed condition. On the other hand, the additional pre-activation for weak role-appropriate competitors in other EEG studies such as Chow, Smith, et al. (2016) does not result in significantly stronger inhibition because weakly activated candidates inhibit their competitors minimally due the sigmoid inhibition function, which is independently motivated

by Chen and Mirman (2012).

However, one problem for this account is that it is not clear how the partial role-sensitivity in pre-activation can be incorporated into the pre-activation mechanisms. If the general role-insensitivity arises from the format mismatch between the contextual cues and the event knowledge, there should be no sensitivity to argument roles at all. The format mismatch account needs to be augmented so that it can allow for the partial role-sensitivity to account for N400 effects found in Ehrenhofer et al..

Symmetric pre-activation: Selective use of cues

Alternatively, the pre-activation of role-appropriate and inappropriate candidates might be matched not because of the format of the input to the pre-activation mechanisms but because of the computations involved in the pre-activation mechanism. Kuperberg (2016) argued within a hierarchical generative framework of language comprehension (Kuperberg & Jaeger, 2016) that argument roles can constrain predictions, but that comprehenders do not use them in cases where using them makes predictions less constraining. For example, when the using unordered pair of arguments (e.g., *waitress*, *customer* in (40a)) results in more constraining predictions with a dominant candidate (*served*) compared to predictions based on the combination of the arguments and their roles (e.g., *customer*-agent, *waitress*-patient in (40a)), comprehenders rely on the unordered pair.

This account also has to explain the three cases where pre-activation measures show sensitivity to argument roles: (i) the speeded cloze task, (ii) N400 amplitudes with delayed presentation of the target items, and (iii) N400 amplitudes with strong role-appropriate competitors in Ehrenhofer et al. (2019). The selective use of cues can in fact explain (iii).

Since there are strong role-appropriate candidates, comprehenders use the full set of cues and predict role-appropriate candidates more than role-inappropriate candidates.

However, it is not straightforward for this hypothesis to account for (i) and (ii). It attributes the strong pre-activation of role-inappropriate candidates to a mechanism that pre-activates candidates based on partial use of cues, but it is not clear why simply adding time between the context and the target verb would change the use of cues, as shown in the Japanese EEG experiment. Even if the use of cues could change over time, the hypothesis still cannot explain why people should pre-activate lexical candidates differently in different paradigms with the same context in the same time course, as I demonstrated in the Japanese speeded cloze experiment.

Furthermore, the selective use of cues has difficulty in accounting for the absence of N400 effects in the Japanese EEG study without delayed presentation of verbs, where only one argument was presented as the context. Given such contexts, it is unlikely that using partial cues (e.g. *bee*) would yield a more dominant candidate than when all the cues are used (e.g. *bee*-patient). Then, Kuperberg (2016)'s account should expect N400 effects for the Japanese stimuli such as *bee*-ACC, which was not the case when the verbs were presented shortly after the contexts.

Given these findings, the format mismatch account augmented by the monitoring mechanism and the partial role-sensitivity seems to be the most reasonable account for the role-sensitivity and insensitivity observed in pre-activation measures. However, the account has to be refined to allow for the partial role-sensitivity to explain the role-sensitivity caused by strong role-appropriate candidates in Ehrenhofer et al. (2019).

7.3 Future directions

This thesis aimed to understand lexical prediction through updating the Simple Activation Model as a working model, and this approach brought new insights into lexical prediction, especially regarding argument roles. While I adopted the model as the starting point, the approach allows for further investigation of lexical prediction by elaborating the model. Below I introduce three directions in which the project can be further developed, in terms of three aspects of the model that could be elaborated: theories of measures, accounts of the inputs to the pre-activation mechanisms, and the structure of the shared currency bridging the shared pre-activation mechanisms and task specific processes.

7.3.1 Theories of measures

While the current study has contrasted multiple measures of pre-activation, the landscape is incomplete. Specifically, eye-tracking studies in both reading and the visual world paradigm are not fully integrated.

Although there are relatively few eye-tracking studies on argument role reversals, these measures have a unique value. Eye movements during reading yield different measures (e.g., skipping rate, first fixation, go-past time) and each measure can tap into different aspects of language processing. The methods to analyze the data (e.g., distributional analyses of reading times: Staub, 2011) and computational models (e.g., E-Z reader model: Pollatsek, Reichle, & Rayner, 2006) further make eye-tracking measures of reading informative about lexical prediction. Fixations in the visual world paradigm allow investigation of dynamic change in pre-activation through shifts in the proportions of fixations. This is especially useful in the studies of argument role reversal, where pre-activation seems to change dynamically, as the

EEG studies showed. Thus the theories of argument roles in lexical prediction, or other cases of lexical prediction, would be augmented by incorporating those measures.

7.3.2 Inputs to the pre-activation mechanisms

This thesis focused on argument role reversals as a test case and it explored whether and how argument roles impact lexical prediction when they are the inputs to the pre-activation mechanisms. This can be further developed in the context of different but highly related phenomena regarding event-related items.

Nieuwland and Van Berkum (2005) presented (42a) or (42b) following a discourse context setting up a situation (41). Surprisingly, there was no contrast in the N400 amplitudes of *tourist* and *suitcase*, suggesting that their pre-activation levels were matched, despite the apparent contrast in predictability.

(41) A tourist wanted to bring his huge suitcase onto the airplane. However, because the suitcase was so heavy, the woman behind the check-in counter decided to charge the tourist extra. In response, the tourist opened his suitcase and threw some stuff out. So now, the suitcase of the resourceful tourist weighed less than the maximum twenty kilos.

(42) a. Next, the woman told the tourist that ...

b. Next, the woman told the suitcase that ...

Relatedly, Metusalem et al. (2012) conducted a similar study and contrasted three types of target items that were (i) predictable given the contexts in the position (*snowman*), (ii) related to the event but not to the predictable item (*jacket*), or (iii) unrelated to either the

event or the predictable item (*towel*). Metusalem et al. found that the event-related item had smaller N400 amplitudes compared to the event-unrelated item.

(43) A huge blizzard ripped through town last night. My kids ended up getting the day off from school. They spent the whole day outside building a big snowman/jacket/towel in the front yard.

These findings are different from argument role reversals in that Nieuwland and Van Berkum (2005) repeated the unpredictable items (*suitcase*) multiple times in the contexts and hence it could have been strongly primed, and that the N400 amplitudes of the event-related items were still larger than those of the predictable items in Metusalem et al. (2012). However, these related findings suggest that argument role reversals might be a part of a broader class of phenomena regarding exaggerated pre-activation for event-related lexical items, where argument role reversals are possibly an extreme case. Elaboration of the Simple Activation Model model (e.g., how event knowledge is probed) can be propelled by situating argument role reversals in these contexts and examining what are specific about arguments and what generally apply to event-related lexical items.

7.3.3 Structure of the shared currency

The Simple Activation Model assumes that lexical (lemma-level) pre-activation is the underlying shared currency that is the output of the pre-activation mechanisms and the common input to models of pre-activation measures. However, the output of the pre-activation mechanism is likely to be represented at multiple levels, since various studies have reported that prediction influences different levels of representations from structures of events to phono-

logical or orthographic representations (See Kuperberg and Jaeger (2016) for review). These different levels of prediction should interact with each other, and each pre-activation measure may reflect a different level or a combination of different levels. Regarding this complexity in the output of pre-activation mechanisms, I made the simplifying assumption that lexical pre-activation is the shared currency. This assumption can be justified if predictions at different levels are closely linked to each other and covary. For example, if the events of a cop arresting a burglar and a cop arresting a smuggler are pre-activated to similar extents given the context (44), the lexical representations (or the semantic representations) of *burglar* and *smuggler* should be pre-activated to similar extents as well.

(44) A cop arrested a ...

However, such strong correlations between pre-activation in different levels of representations must be questioned at least in some cases. In fact, the race process proposed as the underlying mechanism of the (speeded) cloze task can break the correlation between different levels. The race process can be interpreted as a winner-take-all process where one candidate is selected from many candidates and is passed down to the subsequent processes. However, other non-winning candidates are not considered in the subsequent processes, regardless of how closely they lost. Under this formulation, if the race process happens at the event level and if the event of a cop arresting a burglar wins, then the lexical representations of *burglar* should be pre-activated, but the lexical representations of *smuggler* should not be pre-activated, even though the corresponding event of a cop arresting a smuggler was strongly pre-activated. Thus the race processes can break the correlation between the predictions across levels of representations, and what level the race processes are in can have great

consequences on the patterns of pre-activation across levels of representations. Therefore, elaboration of the Simple Activation Model in the structure of the output of the pre-activation mechanism, and how they are reflected in different measures would be informative about the nature of prediction.

7.4 Conclusion

In this thesis, I investigated lexical prediction in human sentence processing, using the Simple Activation Model as a starting point and updating it. Argument role reversals provided insights into the way the model should be updated both in terms of the mechanisms to generate pre-activation and the theories of different experimental measures of pre-activation. While there is not yet a single account that can explain all the patterns across different pre-activation measures, different stimuli, and different timings, the current thesis suggests that role-appropriate and inappropriate lexical candidates are pre-activated similarly, that there is a monitoring process to inhibit role-inappropriate candidates, and that there might be some small advantage for role-appropriate candidate.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the Time Course of Spoken Word Recognition Using Eye Movements: Evidence for Continuous Mapping Models. *Journal of Memory and Language*, *38*(4), 419–439. doi: 10.1006/jmla.1997.2558

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264. doi: 10.1016/S0010-0277(99)00059-1

Altmann, G. T. M., & Mirković, J. (2009). Incrementality and Prediction in Human Sentence Processing. *Cognitive Science*, *33*(4), 583–609. doi: 10.1111/j.1551-6709.2009.01022.x

Andrews, S., & Heathcote, A. (2001). Distinguishing common and task-specific processes in word identification: A matter of some moment? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 514–544. doi: 10.1037/0278-7393.27.2.514

Balota, D. A., Yap, M. J., Cortese, M. J., & Watson, J. M. (2008). Beyond mean response latency: Response time distributional analyses of semantic priming. *Journal of Memory and Language*, *59*(4), 495–523. doi: 10.1016/j.jml.2007.10.004

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*, 1–48. doi: 10.18637/jss.v067.i01

Bicknell, K., Elman, J. L., Hare, M., McRae, K., & Kutas, M. (2010). Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language*, *63*(4), 489–505. doi: 10.1016/j.jml.2010.08.004

Boersma, P., & Weenink, D. (2023). *Praat: Doing Phonetics by Computer.*

Bornkessel-Schlesewsky, I., Kretzschmar, F., Tune, S., Wang, L., Genç, S., Philipp, M., … Schlesewsky, M. (2011). Think globally: Cross-linguistic variation in electrophysiological activity during sentence comprehension. *Brain and Language*, *117*(3), 133–152. doi: 10.1016/j.bandl.2010.09.010

Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. (2017). A Neurocomputational Model of the N400 and the P600 in Language Processing. *Cognitive Science*, *41*(S6), 1318–1352. doi: 10.1111/cogs.12461

Brouwer, H., Fitz, H., & Hoeks, J. (2012). Getting real about Semantic Illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research*, *1446*, 127–143. doi: 10.1016/j.brainres.2012.01.055

Brown, C., & Hagoort, P. (1993). The Processing Nature of the N400: Evidence from Masked Priming. *Journal of Cognitive Neuroscience*, *5*(1), 34–44. doi: 10.1162/jocn.1993.5.1.34

Burnsky, J. (2022). *What did you expect? An Investigation of Lexical Preactivation in Sentence processing* (Unpublished doctoral dissertation). University of Massachusetts Amherst.

Chen, Q., & Mirman, D. (2012). Competition and cooperation among similar representations: Toward a unified account of facilitative and inhibitory effects of lexical neighbors. *Psychological Review*, *119*(2), 417–430. doi: 10.1037/a0027175

Chow, W.-Y., Kurenkov, I., Kraut, B., & Phillips, C. (2015). How predictions change over time: Evidence from an online cloze paradigm. In *The 28th Annual CUNY Conference on Human Sentence Processing.* Los Angeles, CA.

Chow, W.-Y., Lau, E., Wang, S., & Phillips, C. (2018). Wait a second! Delayed impact of argument roles on on-Line verb prediction. *Language, Cognition and Neuroscience*, *33*(7), 803–828. doi: 10.1080/23273798.2018.1427878

Chow, W.-Y., Momma, S., Smith, C., Lau, E., & Phillips, C. (2016). Prediction as memory retrieval: Timing and mechanisms. *Language, Cognition and Neuroscience*, *31*(5), 617–627. doi: 10.1080/23273798.2016.1160135

Chow, W.-Y., Smith, C., Lau, E., & Phillips, C. (2016). A "Bag-of-Arguments" mechanism for initial verb predictions. *Language, Cognition and Neuroscience*, *31*(5), 577–596. doi: 10.1080/23273798.2015.1066832

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. M. (2001). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes*, *16*(5-6), 507–534. doi: 10.1080/01690960143000074

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117–1121. doi: 10.1038/nn1504

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. doi: 10.1016/j.jneumeth.2003.10.009

Ehrenhofer, L., Lau, E., & Phillips, C. (2019). *A Possible Cure for 'N400 Blindness' to Role Reversal Anomalies in Sentence Comprehension* [Submitted].

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, *20*(6), 641–655. doi: 10.1016/S0022-5371(81)90220-6

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, *44*(4), 491–505. doi: 10.1111/j.1469-8986.2007.00531.x

Federmeier, K. D., & Kutas, M. (1999). A Rose by Any Other Name: Long-Term Memory Structure and Sentence Processing. *Journal of Memory and Language*, *41*(4), 469–495. doi: 10.1006/jmla.1999.2660

Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research*, *1146*, 75–84. doi: 10.1016/j.brainres.2006.06.101

Ferretti, T. R., McRae, K., & Hatherell, A. (2001). Integrating Verbs, Situation Schemas, and Thematic Role Concepts. *Journal of Memory and Language*, *44*(4), 516–547. doi:

10.1006/jmla.2000.2728

Friederici, A. D., & Frisch, S. (2000). Verb Argument Structure Processing: The Role of Verb-Specific and Argument-Specific Information. *Journal of Memory and Language*, *43*(3), 476–507. doi: 10.1006/jmla.2000.2709

Hoeks, J. C. J., Stowe, L. A., & Doedens, G. (2004). Seeing words in context: The interaction of lexical and sentence level information during reading. *Cognitive Brain Research*, *19*(1), 59–73. doi: 10.1016/j.cogbrainres.2003.10.022

Indefrey, P., & Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. *Cognition*, *92*(1), 101–144. doi: 10.1016/j.cognition.2002.06.001

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*(1), 133–156. doi: 10.1016/S0749-596X(03)00023-8

Kim, A., & Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Language*, *52*(2), 205–225. doi: 10.1016/j.jml.2004.10.002

Kleiman, G. M. (1980). Sentence frame contexts and lexical decisions: Sentence-acceptability and word-relatedness effects. *Memory & Cognition*, *8*(4), 336–344. doi: 10.3758/BF03198273

Kolk, H. H. J., Chwilla, D. J., van Herten, M., & Oor, P. J. W. (2003). Structure and limited capacity in verbal working memory: A study with event-related potentials. *Brain and Language*, *85*(1), 1–36. doi: 10.1016/S0093-934X(02)00548-5

Kukona, A., Fang, S.-Y., Aicher, K. A., Chen, H., & Magnuson, J. S. (2011). The time course of anticipatory constraint integration. *Cognition*, *119*(1), 23–42. doi: 10.1016/j.cognition.2010.12.002

Kuperberg, G. R. (2016). Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. *Language, Cognition and Neuroscience*, *31*(5), 602–616. doi: 10.1080/23273798.2015.1130233

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59. doi: 10.1080/23273798.2015.1102299

Kuperberg, G. R., Sitnikova, T., Caplan, D., & Holcomb, P. J. (2003). Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Cognitive Brain Research*, *17*(1), 117–129. doi: 10.1016/S0926-6410(03)00086-7

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, *4*(12), 463–470. doi: 10.1016/S1364-6613(00)01560-6

Kutas, M., & Federmeier, K. D. (2011). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology*, *62*(1), 621–647. doi: 10.1146/annurev.psych.093008.131123

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163. doi: 10.1038/307161a0

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*, 1–26. doi: 10.18637/jss.v082.i13

Lau, E. F., Holcomb, P. J., & Kuperberg, G. R. (2013). Dissociating N400 Effects of Prediction from Association in Single-word Contexts. *Journal of Cognitive Neuroscience*, *25*(3), 484–502. doi: 10.1162/jocn$_{a0}$0328

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400. *Nature Reviews Neuroscience*, *9*(12), 920–933. doi: 10.1038/nrn2532

Lee, E.-K., Howitt, K., Dixon, L., Ness, T., Nakamura, M., Muller, H., & Phillips, C. (2023). *A shared underlying mechanism in children's and adults' predictive processing: Evidence from a speeded cloze paradigm.*

Lee, E.-K., Nakamura, M., Muller, H., & Phillips, C. (2022). Argument roles and verb expectations: Narrowing a comprehension-production contrast. In *35th Annual Conference on Human Sentence Processing.* Santa Cruz, CA.

Lee, E.-K., & Phillips, C. (2023). Top-down goals modulate the use of argument roles in prediction: Evidence from ERPs. In *36th Annual Conference on Human Sentence Processing.* Pittsburgh, PA.

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition*, *14*(1), 41–104. doi: 10.1016/0010-0277(83)90026-4

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* Cambridge, MA, US: The MIT Press.

Li, J., & Ettinger, A. (2023). Heuristic interpretation as rational inference: A computational model of the N400 and P600 in language processing. *Cognition*, *233*, 105359. doi: 10.1016/j.cognition.2022.105359

Liao, C.-H., Lau, E., & Chow, W.-Y. (2022). Towards a processing model for argument-verb computations in online sentence comprehension. *Journal of Memory and Language*, *126*, 104350. doi: 10.1016/j.jml.2022.104350

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, *8*.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing Spoken Words: The Neighborhood Activation Model. *Ear and hearing*, *19*(1), 1–36.

Luce, R. D. (1959). *Individual choice behavior.* Oxford, England: John Wiley.

Luck, S. J. (2021). *Applied Event-Related Potential Data Analysis.* LibreTexts. doi: 10.18115/D5QG92

Marslen-Wilson, W. (1990). Activation, Competition, and Frequency in Lexical Access. In *Cognitive models of speech processing: Psycholinguistic and computational perspectives*

(pp. 148–172).

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86. doi: 10.1016/0010-0285(86)90015-0

Metusalem, R., Kutas, M., Urbach, T. P., Hare, M., McRae, K., & Elman, J. L. (2012). Generalized event knowledge activation during online sentence comprehension. *Journal of Memory and Language*, *66*(4), 545–567. doi: 10.1016/j.jml.2012.01.001

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*, 227–234. doi: 10.1037/h0031564

Momma, S. M. (2016). *Parsing, Generation and Grammar* (Unpublished doctoral dissertation). University of Maryland, College Park.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, *76*(2), 165–178. doi: 10.1037/h0027366

Ness, T., & Meltzer-Asscher, A. (2018). Predictive Pre-updating and Working Memory Capacity: Evidence from Event-related Potentials. *Journal of Cognitive Neuroscience*, *30*(12), 1916–1938. doi: 10.1162/jocn$_{a0}$1322

Ness, T., & Meltzer-Asscher, A. (2021a). From pre-activation to pre-updating: A threshold mechanism for commitment to strong predictions. *Psychophysiology*, *58*(5), e13797. doi: 10.1111/psyp.13797

Ness, T., & Meltzer-Asscher, A. (2021b). Love thy neighbor: Facilitation and inhibition in the competition between parallel predictions. *Cognition*, *207*, 104509. doi: 10.1016/j.cognition.2020.104509

Nieuwland, M. S., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., … Von Grebmer Zu Wolfsthurn, S. (2020). Dissociable effects of prediction and integration during language comprehension: Evidence from a large-scale study using brain potentials. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1791), 20180522. doi: 10.1098/rstb.2018.0522

Nieuwland, M. S., & Kuperberg, G. R. (2008). When the Truth Is Not Too Hard to Handle: An Event-Related Potential Study on the Pragmatics of Negation. *Psychological Science*, *19*(12), 1213–1218. doi: 10.1111/j.1467-9280.2008.02226.x

Nieuwland, M. S., & Van Berkum, J. J. A. (2005). Testing the limits of the semantic illusion phenomenon: ERPs reveal temporary semantic change deafness in discourse comprehension. *Cognitive Brain Research*, *24*(3), 691–701. doi: 10.1016/j.cogbrainres.2005.04.003

Nieuwland, M. S., & Van Berkum, J. J. A. (2006). When Peanuts Fall in Love: N400 Evidence for the Power of Discourse. *Journal of Cognitive Neuroscience*, *18*(7), 1098–1111. doi: 10.1162/jocn.2006.18.7.1098

Oishi, H., & Tsutomu, S. (2010). *The integration of outputs from syntactic and semantic/thematic processing: "semantic P600" effects in Japanese sentence processing.*

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *2011*, 1:1–1:9. doi: 10.1155/2011/156869

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1), 8–13. doi: 10.1016/j.jneumeth.2006.11.017

Perea, M., & Rosa, E. (2002). The effects of associative and semantic priming in the lexical decision task. *Psychological Research*, *66*(3), 180–194. doi: 10.1007/s00426-002-0086-5

Pollatsek, A., Reichle, E. D., & Rayner, K. (2006). Tests of the E-Z Reader model: Exploring the interface between cognition and eye-movement control. *Cognitive Psychology*, *52*(1), 1–56. doi: 10.1016/j.cogpsych.2005.06.001

R Core Team. (2020). *R: A language and environment for statistical computing* [Manual]. Vienna, Austria.

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modeling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, *2*(9), 693–705. doi: 10.1038/s41562-018-0406-4

Rabs, E., Delogu, F., Drenhaus, H., & Crocker, M. W. (2022). Situational expectancy or association? The influence of event knowledge on the N400. *Language, Cognition and Neuroscience*, *37*(6), 766–784. doi: 10.1080/23273798.2021.2022171

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC Nonword Database. *The Quarterly Journal of Experimental Psychology Section A*, *55*(4), 1339–1362. doi: 10.1080/02724980244000099

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin*, *86*, 446–461. doi: 10.1037/0033-2909.86.3.446

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A Diffusion Model Account of the Lexical Decision Task. *Psychological review*, *111*(1), 159–182. doi: 10.1037/0033-295X.111.1.159

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Sciences*, *20*(4), 260–281. doi: 10.1016/j.tics.2016.01.007

Reichle, E. D., Pollatsek, A., & Rayner, K. (2012). Using E-Z Reader to simulate eye movements in nonreading tasks: A unified framework for understanding the eye–mind link. *Psychological Review*, *119*(1), 155–185. doi: 10.1037/a0026473

Roux, F., Armstrong, B. C., & Carreiras, M. (2017). Chronset: An automated tool for detecting speech onset. *Behavior Research Methods*, *49*(5), 1864–1881. doi: 10.3758/s13428-016-0830-1

Seidenberg, M. S., Waters, G. S., Sanders, M., & Langer, P. (1984). Pre- and postlexical loci of contextual effects on word recognition. *Memory & Cognition*, *12*(4), 315–328. doi: 10.3758/BF03198291

Smith, N., & Levy, R. (2011). Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. *Proceedings of the Annual*

Meeting of the Cognitive Science Society, *33*(33).

Stanovich, K. E., & West, R. F. (1983). On priming by a sentence context. *Journal of Experimental Psychology: General*, *112*, 1–36. doi: 10.1037/0096-3445.112.1.1

Staub, A. (2011). The effect of lexical predictability on distributions of eye fixation durations. *Psychonomic Bulletin & Review*, *18*(2), 371–376. doi: 10.3758/s13423-010-0046-9

Staub, A. (2015). The Effect of Lexical Predictability on Eye Movements in Reading: Critical Review and Theoretical Interpretation. *Language and Linguistics Compass*, *9*(8), 311–327. doi: 10.1111/lnc3.12151

Staub, A., Grant, M., Astheimer, L., & Cohen, A. (2015). The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language*, *82*, 1–17. doi: 10.1016/j.jml.2015.02.004

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634. doi: 10.1126/science.7777863

Taylor, W. L. (1953). "Cloze Procedure": A New Tool for Measuring Readability. *Journalism Bulletin*, *30*(4), 415–433. doi: 10.1177/107769905303000401

van Herten, M., Kolk, H. H. J., & Chwilla, D. J. (2005). An ERP study of P600 effects elicited by semantic anomalies. *Cognitive Brain Research*, *22*(2), 241–255. doi: 10.1016/j.cogbrainres.2004.09.002

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*(2), 176–190. doi: 10.1016/j.ijpsycho.2011.09.015

Venhuizen, N. J., Crocker, M. W., & Brouwer, H. (2019). Expectation-based Comprehension: Modeling the Interaction of World Knowledge and Linguistic Experience. *Discourse Processes*, *56*(3), 229–255. doi: 10.1080/0163853X.2018.1448677

Wlotko, E. W., & Federmeier, K. D. (2012). So that's what you meant! Event-related potentials reveal multiple aspects of context use during construction of message-level meaning. *NeuroImage*, *62*(1), 356–366. doi: 10.1016/j.neuroimage.2012.04.054

Xiang, M., & Kuperberg, G. (2015). Reversing expectations during discourse comprehension. *Language, Cognition and Neuroscience*, *30*(6), 648–672. doi: 10.1080/23273798.2014.995679

Zehr, J., & Schwarz, F. (2018). PennController for Internet Based Experiments (IBEX). doi: 10.17605/OSF.IO/MD832